THIS WEEK

With transparency comes trust

International development experts say that the Millennium Villages Project's claims of progress should be interpreted with caution.

There is an intuitive appeal to the Millennium Villages international development project — the brainchild of economist Jeffrey Sacks from Columbia University in New York that aims to help lift villages in 14 sites across Africa from poverty. The initiative takes a broad approach and aims to tackle the root causes of poverty and ill health together, unlike most aid projects, which focus on just one area.

Improvements on the ground seem impressive. In Mwandama, Malawi, crop yields shot up after the project gave farmers free fertilizer and improved varieties of maize seeds. Any food not used immediately is stored in a purpose-built warehouse and sold later in the year, earning the villagers an income. With the project's support, many villagers are now able to grow mango trees and sweet potatoes to improve food security, diversify their diet and earn extra money.

Freed from the daily struggle to fill their bellies and armed with better access to health care and education, the villagers are now setting up cooperative business ventures and investments so that they can support themselves without the project's assistance.

The villagers of Mwandama, and probably the inhabitants of the other selected villages, are clearly better off than they were six years ago, when the project started. And given the economic crisis, the project has done well to attract the funds that it has.

But prominent international development researchers and experts have taken issue with some of the project's claims of progress (see page 158) — most recently for declines in child mortality (P. M. Pronyk *et al. Lancet* http://dx.doi.org/10.1016/s0140-6736(12)60207-4; 2012). Their main concerns, they say, are weaknesses in the project's design and data analysis, as well as a lack of transparency over the raw data and project costs.

Michael Clemens, a development researcher at the Center for Global Development in Washington DC, has done some of the most detailed independent analyses of the project. He questions some of the key findings in the latest paper, in part because of inadequate baseline data for child mortality in the control villages.

The project started to monitor control villages only after three years had passed, when independent development researchers argued that it was necessary. So, in the absence of actual data for that period, the study asked female villagers to recall child deaths over the previous three years.

The paper does acknowledge that this method is unreliable and can underestimate child deaths. So how reliable are comparisons made against such data? Clemens says that alarm bells should have started ringing when the verbal reports suggested that the child death rate in the control villages rose over the study period, contradicting continent-wide trends of falling child mortality. This would have resulted in an overestimation of the difference in child mortality between the two sites, Clemens argues.

Furthermore, he says that the study's statistical analysis fails to show that the annual rate of decline in the project villages is triple that of national rural trends, as claimed. And, he says, the data set used for national child mortality was a misleading comparison because it includes the period 2000–06, before the villages project started, when the national child death rate was falling more slowly than in 2006–10. This, he says, made national rural trends seem lower than they actually are, thereby inflating the improvements in the project villages.

Nature put these criticisms to the organizers of the Millennium Villages Project and received lengthy written and verbal responses. Not all of these helped to clarify the situation. When asked, for example, why the claimed improvements in child mortality were placed in such a prominent position in the paper despite being statistically insignificant, the organizers replied that they were there for "illustrative" purposes.

"The project's approach has potential, but little can be said for sure about its true impact."

The Millennium Villages Project has problems beyond the analysis of data. The organizers have been reluctant to publish a full breakdown of costs — making it impossible for those not privy to the information to verify their cost–benefit analysis, which is crucial in development policy because spending is under great scrutiny. The project also seems to lack a coherent policy on when and how it will make data available to independent researchers.

Clemens and others are right to ask that the project make this information available. Greater transparency is essential to build trust and credibility. The project's approach has potential, but little can be said for sure yet about its true impact. The latest initiative of the Millennium Villages Project, in Ghana and funded by the UK government, is a welcome step in the right direction. It builds in independent scrutiny from the start, and has been open and transparent about its costs. All future projects should follow this model. ■

Misplaced protest

Rothamsted's genetically engineered wheat should be allowed to grow.

Plant scientists at Rothamsted Research, a complex of buildings and fields in Hertfordshire, UK, that prides itself on being the longest-running agricultural research station in the world, have spent years preparing for their latest experiment — which will attempt to prove the usefulness of a genetically modified (GM) wheat that emits an aphid alarm pheromone, potentially reducing aphid infestation.

Yet instead of looking forward to watching their crop grow, the Rothamsted scientists are nervously counting the days until 27 May,

when protesters against GM crops have promised to turn up in force and destroy the experimental plots.

The protest group, it must be acknowledged, has a great name — Take the Flour Back. And it no doubt believes that it has the sympathy of the public. The reputation of GM crops and food in Britain, and in much of mainland Europe, has yet to recover from the battering it took in the late 1990s. In Germany, the routine destruction of crops by protesters has meant that scientists there simply don't bother to conduct GM experiments any more.

The Rothamsted scientists have also attempted to win over the public, with a media campaign that explains what they are trying to do and why. After the protesters announced their plans to "decontaminate" the research site, the scientists tried to engage with their opponents, and pleaded with them to "reconsider before it is too late, and before years of work to which we have devoted our lives are destroyed forever". The researchers say that in this case they are the true environmentalists. The modified crop, if it works, would lower the demand for environmentally damaging insecticides.

As *Nature* went to press, the stalemate continued. The GM crop at Rothamsted remains, but so does the intention of the protesters to destroy it.

There are very real consequences to this kind of protest. German chemical giant BASF this year announced that it would move its transgenic plant operations from Europe to the United States, in part because of the perception of continuing widespread opposition to GM crops in Europe. And although farmers in other parts of the world have taken to GM crops with gusto, Europe, with some exceptions, misses out. Evidence suggests that it is missing a lot. The adoption of herbicide-resistant oilseed rape has reduced the use of herbicides by farmers in North America, and also reduced tillage, which has its own environmental benefits. The adoption of pest-resistant GM cotton has lowered the use of pesticides. Nevertheless, the reasons for the hostility towards genetic modification in Europe are clear. Justifiable unease over the way in which GM-led business models would hand entire food chains to large agrochemical companies found a popular proxy in less-realistic concerns over the possible health impacts of the new technology.

But with the world's population now at 7 billion and counting, the rejection of genetic modification of crops on such spurious scientific grounds now threatens the environment it claims to protect. To feed a population likely to top 9 billion in 2100, we are going to need to change the way we grow our food. Harking back to old-fashioned methods and talking up organic farming will not do it. Genetic modification alone will not do it, but it could be a crucial tool and one that it is foolish to oppose on sentimental or ideological grounds.

"To destroy experiments before the outstanding questions can be answered is more than local vandalism, it is recklessness on a global scale."

This will not convince diehard opponents, of course, just as pleas for the value of scientific research failed to sway the criminal faction of the animal-rights movement. But, just as it proved with animal rights, it is far from clear that GM protesters, however many turn up at Rothamsted in a fortnight, truly attract public support.

GM crops could significantly reduce the use of pesticides, herbicides and fertilizers, and provide greater tolerance to a more extreme climate. True, we are still in the early stages of this technology. And there are some legitimate concerns, such as possible leakage of GM material into the local environment. But to destroy experiments such as the one at Rothamsted before the outstanding questions can be answered is more than local vandalism, it is recklessness on a global scale. ■

Price of freedom

The latest mission to Jupiter highlights the benefits and pitfalls of collaboration.

It is a long trip to the outer reaches of the Solar System. Planetary scientists who are eager to explore Jupiter and the planets beyond tend to plan their experiments not in terms of years, but generations. And so it is with some rejoicing, and also relief, that they have another mission on the books.

Last week, the European Space Agency (ESA) announced that it had selected the Jupiter Icy moons Explorer, or JUICE, a solar-powered behemoth that, at 4.8 tonnes, would be the heaviest interplanetary probe ever flown by Europe. It would launch in 2022 and arrive at Jupiter almost eight years later. After a few flybys of Jupiter's moons Callisto and Europa, in 2032 the probe would settle into orbit around its primary target, the moon Ganymede, for at least a year of science. Ganymede's main mystery is its enigmatic magnetic field, the only moon in the Solar System to have one. But, like Europa, Ganymede also has a subsurface ocean — although one that is less enticing to astrobiologists because it is likely to be isolated, sandwiched between thick layers of ice that prevent interesting chemical interactions with the surface and the deep rocky mantle.

Still, JUICE came top in a competition that sent two other prospective European missions packing. One was an X-ray telescope that would have imaged objects such as black holes with greater precision and sensitivity than ever before. Another was a set of satellites that, flying in formation, would have sensed tiny ripples in the fabric of space caused by violent events such as black-hole mergers — thereby opening up a whole new field: observational gravitational-wave astronomy.

Neither mission was a dud scientifically; quite the opposite. The gravitational-wave mission, in particular, is viewed as representing a scientific revolution in the making. These missions failed in the competition because they were expensive, and were likely to bust ESA's budget of €1 billion (US$1.3 billion). And the reason ESA could not afford them was because both were originally designed as joint missions with the United States. When NASA pulled out, each mission tried to reduce its scope and lower its price tag, but that proved too difficult.

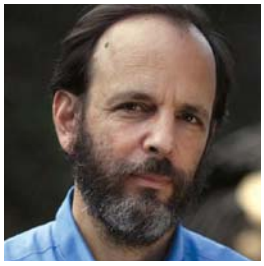
JUICE was also once married to a NASA mission, but in a more modern arrangement. The ESA mission would have had its own satellite and rocket launcher, as would NASA, which would have sent an orbiter devoted to studying Europa. When the budgetary rug was pulled out from under NASA's Europa orbiter, JUICE was in much better shape, politically and financially.

The lessons here would seem to be perverse: eschew tight collaborations and you will be rewarded for your independence. Avoid working with foreign agencies and you will be better off in the long run.

That might be true, but only from the perspective of a scientist interested in Ganymede — and only Ganymede. Without NASA involvement, plenty of European scientists would be lost. And had the two missions launched as a loose partnership, there would have been several ways in which the sum of the two missions was greater than its parts. For example, tracking the magnetosphere of the Jovian system using two probes makes a far better map than using just one.

The bigger point, however, is that the frontiers of science in many fields are reaching the stage — or price tag — at which no single country can go it alone. Just ask scientists who worked on completing the Human Genome Project, or building the Large Hadron Collider near Geneva, Switzerland. Of late, space scientists at NASA and ESA have no such project to hold up as an example. In addition to the X-ray and gravitational-wave observatories, other transatlantic partnerships have evaporated, including ones to study dark energy and to return samples from Mars. If a mission to the king of the planets is a cause for rejoicing, then the fact that it is so singular may be a cause for alarm. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv



Beware the creeping cracks of bias

Evidence is mounting that research is riddled with systematic errors. Left unchecked, this could erode public trust, warns **Daniel Sarewitz**.

Alarming cracks are starting to penetrate deep into the scientific edifice. They threaten the status of science and its value to society. And they cannot be blamed on the usual suspects — inadequate funding, misconduct, political interference, an illiterate public. Their cause is bias, and the threat they pose goes to the heart of research.

Bias is an inescapable element of research, especially in fields such as biomedicine that strive to isolate cause–effect relations in complex systems in which relevant variables and phenomena can never be fully identified or characterized. Yet if biases were random, then multiple studies ought to converge on truth. Evidence is mounting that biases are not random. A Comment in *Nature* in March reported that researchers at Amgen were able to confirm the results of only six of 53 'landmark studies' in preclinical cancer research (C. G. Begley & L. M. Ellis *Nature* **483,** 531–533; 2012). For more than a decade, and with increasing frequency, scientists and journalists have pointed out similar problems.

Early signs of trouble were appearing by the mid-1990s, when researchers began to document systematic positive bias in clinical trials funded by the pharmaceutical industry. Initially these biases seemed easy to address, and in some ways they offered psychological comfort. The problem, after all, was not with science, but with the poison of the profit motive. It could be countered with strict requirements to disclose conflicts of interest and to report all clinical trials.

Yet closer examination showed that the trouble ran deeper. Science's internal controls on bias were failing, and bias and error were trending in the same direction — towards the pervasive over-selection and over-reporting of false positive results. The problem was most provocatively asserted in a now-famous 2005 paper by John Ioannidis, currently at Stanford University in California: 'Why Most Published Research Findings Are False' (J. P. A. Ioannidis *PLoS Med.* **2,** e124; 2005). Evidence of systematic positive bias was turning up in research ranging from basic to clinical, and on subjects ranging from genetic disease markers to testing of traditional Chinese medical practices.

How can we explain such pervasive bias? Like a magnetic field that pulls iron filings into alignment, a powerful cultural belief is aligning multiple sources of scientific bias in the same direction. The belief is that progress in science means the continual production of positive findings. All involved benefit from positive results, and from the appearance of progress. Scientists are rewarded both intellectually and professionally, science administrators are empowered and the public desire for a better world is answered. The lack of incentives to report negative results, replicate experiments or recognize inconsistencies, ambiguities and uncertainties is widely appreciated — but the necessary cultural change is incredibly difficult to achieve.

A BIASED SCIENTIFIC RESULT IS NO DIFFERENT FROM A USELESS ONE.

Researchers seek to reduce bias through tightly controlled experimental investigations. In doing so, however, they are also moving farther away from the real-world complexity in which scientific results must be applied to solve problems. The consequences of this strategy have become acutely apparent in mouse-model research. The technology to produce unlimited numbers of identical transgenic mice attracts legions of researchers and abundant funding because it allows for controlled, replicable experiments and rigorous hypothesis-testing — the canonical tenets of 'scientific excellence'. But the findings of such research often turn out to be invalid when applied to humans.

A biased scientific result is no different from a useless one. Neither can be turned into a real-world application. So it is not surprising that the cracks in the edifice are showing up first in the biomedical realm, because research results are constantly put to the practical test of improving human health. Nor is it surprising, even if it is painfully ironic, that some of the most troubling research to document these problems has come from industry, precisely because industry's profits depend on the results of basic biomedical science to help guide drug-development choices.

Scientists rightly extol the capacity of research to self-correct. But the lesson coming from biomedicine is that this self-correction depends not just on competition between researchers, but also on the close ties between science and its application that allow society to push back against biased and useless results.

It would therefore be naive to believe that systematic error is a problem for biomedicine alone. It is likely to be prevalent in any field that seeks to predict the behaviour of complex systems — economics, ecology, environmental science, epidemiology and so on. The cracks will be there, they are just harder to spot because it is harder to test research results through direct technological applications (such as drugs) and straightforward indicators of desired outcomes (such as reduced morbidity and mortality).

Nothing will corrode public trust more than a creeping awareness that scientists are unable to live up to the standards that they have set for themselves. Useful steps to deal with this threat may range from reducing the hype from universities and journals about specific projects, to strengthening collaborations between those involved in fundamental research and those who will put the results to use in the real world. There are no easy solutions. The first step is to face up to the problem — before the cracks undermine the very foundations of science. ■

➲ NATURE.COM
Discuss this article online at:
go.nature.com/bo2srq

Daniel Sarewitz *is co-director of the Consortium for Science, Policy and Outcomes at Arizona State University, and is based in Washington DC.*
e-mail: dsarewitz@gmail.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

NEUROSCIENCE

BOLD strides in brain imaging

To visualize brain activity, neuroscientists use functional magnetic resonance imaging (fMRI) to measure blood oxygen levels, known as BOLD signals, which are considered a proxy for cellular activity. However, it has been unclear which types of brain cell contribute to these signals.

Fritjof Helmchen and his colleagues at the University of Zurich in Switzerland have developed a method that tracks the activity of neurons and glial cells — support cells that might also contribute indirectly to neurotransmission — during an fMRI scan. They found that activation of both cell types correlates with BOLD signals.

The team used an optical fibre to record the activity of dye-loaded brain cells that fluoresce when calcium enters them — an indication of cell activation. This composite method will help scientists to interpret BOLD signals, the authors say.

Nature Methods **http://dx.doi.org/10.1038/nmeth.2013 (2012)**

TONY WALTHAM/ROBERT HARDING

CLIMATE SCIENCE

A check on speeding glaciers

Analysis of a decade-long record of Greenland's glaciers suggests that the ice sheets are not accelerating towards the ocean as much as previously forecast.

Earlier work on a small number of glaciers had uncovered large increases in speed. Using satellite radar data to calculate the movements of more than 200 of the island's ocean-terminating glaciers between 2000 and 2010, Twila Moon at the University of Washington, Seattle, and her colleagues found a range of accelerations and decelerations, with an overall acceleration. Glaciers in the northwest and southeast — where approximately 80% of ice loss occurs — accelerated by about 30% over the ten-year period, whereas glaciers elsewhere exhibited a generally steady flow.

Glacial melting can lead to an increase in sea level. However, Moon and colleagues' data suggest that Greenland's current glacial acceleration is unlikely to produce the previously forecast worst-case scenario of a 0.5-metre sea-level rise by 2100.

Science 336, **576–578 (2012)**

PHYSIOLOGY

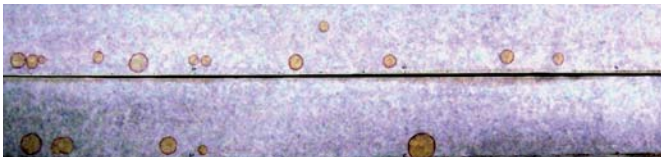
Bladder under circadian control

Most adults produce less urine at night than during the day, and store more of what is made, thanks to the circadian regulation of daily urination patterns.

Hitoshi Okamura and Osamu Ogawa at Kyoto University in Japan and their colleagues developed a machine that measures the urine discharges of mice, as stains on paper (**pictured**), over time. They focused on a protein, connexin43, which increases the frequency of urination by making the bladder muscles more sensitive to neural signals. They found that connexin43 levels peaked during the night when the nocturnal creatures were active. Mice without the circadian clock gene *Cry* produced less connexin43 during the night than did normal mice, and did not show daily rhythms in urination patterns. Another clock gene, *Rev-erbα*, regulates connexin43 expression.

The authors suggest that other genes related to bladder-muscle contraction and daily cycles might also have a role in staving off night-time trips to the toilet.

Nature Commun. **http://dx.doi.org/10.1038/ncomms1812 (2012)**



GEOCHEMISTRY

North Sea starved of oxygen

Summer oxygen levels are declining in some parts of the North Sea, probably because of ocean warming and the decay of photosynthetic blooms that form as a result of nutrient influx.

Bastien Queste at the University of East Anglia in Norwich, UK, and his team compared the results of an oceanographic field survey conducted in August 2010 with twentieth-century records of

oxygen concentrations in the North Sea. They found that the intensity of seasonal oxygen depletion in highly stratified regions — where there is little mixing between layers of water of different temperature — has increased markedly since 1990.

In 2010, dissolved oxygen in the central North Sea and in an area known as the Oyster Grounds near the Dutch coast came close to ecologically critical values that, if reached, would require management action under the European Union's Water Framework Directive, the team reports.

Biogeochemistry **http://dx.doi.org/10.1007/s10533-012-9729-9 (2012)**

EVOLUTION

Gene duplication for bigger brains

DNA-duplication errors that upped the number of copies of a gene may have catalysed the evolution of complex brains in early humans.

The gene *SRGAP2* is expressed during development of the brain's neocortex — a region involved in cognition. Evan Eichler at the University of Washington in Seattle and his team report that humans have four different versions of *SRGAP2*, as did Neanderthals, whereas other primates have just one. The group estimates that successive duplications of *SRGAP2* occurred between 3.4 million and 1 million years ago, as *Homo* species evolved.

Meanwhile, Franck Polleux at the Scripps Research Institute in La Jolla, California, and his team show that one of the newer versions of the gene, *SRGAP2C*, blocks the activity of the ancestral *SRGAP2* when it is artificially expressed in the brains of mice. Mouse neurons expressing *SRGAP2C* develop features of human neurons, such as a denser array of projections called dendritic spines that forge connections with neighbouring neurons. The cells also migrated across the developing brain faster than normal mouse neurons.

The authors suggest that these changes, driven by the emergence of *SRGAP2C*, could have occurred in early humans, who had much larger brains than their ancestors.

Cell **http://dx.doi.org/10.1016/j.cell.2012.03.033; http://dx.doi.org/10.1016/j.cell.2012.03.034 (2012)**

For a longer story on this research, see http://go.nature.com/osz4hk

NANOBIOTECHNOLOGY

Radio remote control of genes

Externally applied radio waves can be used to switch on a modified gene in a mouse, thanks to radiation-absorbing nanoparticles injected into the animal. The technique could enable researchers to activate cells and genes non-invasively.

Jeffrey Friedman at the Rockefeller University in New York and his team coated iron oxide nanoparticles with antibodies so that they bound to a cell-surface protein complex, TRPV1, that admits calcium ions to the cell at a temperature of 42 °C. The researchers used radio waves to heat the nanoparticles, which, in turn, heated TRPV1. Calcium entering the cell activated the gene for an insulin precursor, which had been modified to contain a calcium-sensitive regulatory region.

In live mice, 30 minutes of radio-wave exposure boosted insulin and lowered blood sugar levels.

Science **336, 604–608 (2012)**

BIOPHYSICS

High-throughput cell stretcher

A chip on which cells flow through tiny channels can be used to measure the size and deformability of individual cells at a rate of 2,000 per second — several orders of magnitude faster than existing methods. The chip could be used to detect cancer cells, which are more deformable than healthy cells.

Dino Di Carlo and his team at the University of California, Los Angeles, developed the microfluidic device, which suspends cells single-file in a liquid, stretches them, and then uses automated image analysis to measure their size and rigidity. The team detected cancerous cells in samples from patients with a sensitivity of 91% and a specificity of 86%. The researchers were also able to classify stem cells on the basis of their deformability.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1200107109 (2012)**

COMMUNITY CHOICE

The most viewed papers in science

DRUG DELIVERY

On-demand drug release

HIGHLY READ on pubs.acs.org in April

Drug-carrying nanoparticles that shrink and release their payload when irradiated with ultraviolet (UV) light could offer a way to get drugs deep into tissues and to unleash them on demand. This could be a valuable therapeutic tool for diseases such as cancer.

Currently available drug-delivering nanoparticles are at least 100 nanometres in diameter, which makes it difficult for them to squeeze into tumours. Daniel Kohane at the Children's Hospital Boston in Massachusetts and his colleagues made their particles out of organic molecules that switch conformation when hit with UV light. The nanoparticles were able to carry a number of drugs, including several used in cancer treatment, and shrank from roughly 150 to 40 nanometres under UV light. Irradiated particles released their drug cargo at a higher rate and diffused farther through both a collagen gel and corneal tissue than those not exposed to UV light.

J. Am. Chem. Soc. **http://dx.doi.org/10.1021/ja211888a (2012)**

EVOLUTION

Cheating cuts offspring fitness

'Monogamous' female birds often produce young with another partner. This was presumed to yield offspring fitter than those produced with the paired partner, but a study of song sparrows suggests that 'cheating' comes with no evolutionary reproductive benefit.

Jane Reid at the University of Aberdeen, UK, and her team analysed 17 years' worth of genetic parentage data

REBECCA SARDELL

from a small population of song sparrows (*Melospiza melodia*; nestlings **pictured**) on Canada's Mandarte Island. They compared the lifetime reproductive success of half siblings with the same mother and found that young sired outside of monogamy were less reproductively fit than their half-siblings, producing on average 40% fewer offspring and 30% fewer grand-offspring.

The researchers suggest that there may be indirect selection against, not for, cheating in song sparrows.

Am. Nat. **http://dx.doi.org/10.1086/665665 (2012)**

NATURE.COM

For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

Nuclear-free Japan

Japan's last operating nuclear reactor was switched off on 5 May, leaving the country entirely without nuclear power. The reactor, at the Tomari nuclear power plant in Hokkaido, was taken offline for routine maintenance. None of the reactors that were closed down after an earthquake and tsunami struck the Fukushima Daiichi plant last March has yet reopened. It is not clear when or if any of Japan's 50 functional reactors will go back online; some have passed official safety tests, but face suspicion from local residents. See go.nature.com/dld2nz for more.

Fracking rules

The US Department of the Interior's Bureau of Land Management released a draft on 4 May of rules that would require companies to disclose the chemicals they use in hydraulic fracturing or 'fracking', which involves pumping fluid into rocks to release natural gas and oil. The technique has provoked public protest, in part because of fears that chemicals used in the process could pollute ground water. See go.nature.com/ozuqd2 for more.

Failed drugs review

A cache of at least 24 drugs abandoned during development by three major pharmaceutical companies will become available to scientists seeking new therapeutic uses for them under a US$20-million competitive grants programme, the US National Institutes of Health (NIH) announced on 3 May. The pilot programme, with Pfizer, AstraZeneca and Eli Lilly, is running through the NIH's National Center for Advancing Translational Sciences, which was established this year. A similar collaboration, between Astrazeneca and the UK Medical Research Council, was launched last December. See go.nature.com/o4gjtl for more.

J. GREENBERG/ALAMY

Dinosaur hall set for revamp

The Smithsonian Institution National Museum of Natural History in Washington DC has announced a US$45-million overhaul of its dinosaur hall, mostly paid for by billionaire philanthropist David Koch. Koch is providing $35 million for the revamped hall, which will be named after him. Construction is slated to begin in 2014. The businessman and his brother Charles are controversial figures: they have provided large amounts of money to research projects, but have also donated to campaigns that combat climate science, such as those of the Heartland Institute, a libertarian think tank in Chicago, Illinois.

South Korea carbon

Building on an ambitious effort to invest in green technologies and reduce reliance on imported fossil fuels, South Korea has become the first Asian nation to formally adopt a cap-and-trade programme to reduce greenhouse-gas emissions. The programme, approved on 2 May, starts in 2015 and would cover around 60% of national emissions. The country has committed to reducing its 2020 'business-as-usual' projected emissions by 30%. See go.nature.com/iouadq for more.

BUSINESS

Bio-pharming

The US Food and Drug Administration approved on 2 May its first biological drug produced in a genetically engineered plant cell. The drug, Elelyso (taliglucerase alfa), is made by biotech company Protalix Biotherapeutics based in Carmiel, Israel, to treat Gaucher's disease, a hereditary enzyme-deficiency disorder. See page 160 for more.

Drug-marketing fine

US pharmaceutical company Abbott Laboratories will pay US$1.6 billion in connection with its illegal marketing of an anti-seizure drug, Depakote (divalproex sodium). The company, which is headquartered in Abbott Park, Illinois, and the US Department of Justice announced the settlement on 7 May, ending a four-year investigation. Abbott marketed the medication for uses not approved by the US Food and Drug Administration, including schizophrenia and agitated dementia.

PEOPLE

Physicist sentenced

French-Algerian physicist Adlène Hicheur has been sentenced to four years in prison — and a further one-year suspended sentence — after being found guilty of plotting with al-Qaeda's North African branch to carry out terror attacks on French

soil. Supporters said that his sentencing on 4 May was a miscarriage of justice. Hicheur has been in custody since he was detained in 2009, when he was a postdoc in high-energy physics at the Swiss Federal Institute of Technology in Lausanne. With time already spent in prison, and term reductions available under France's judicial system, he is likely to be released soon. See go.nature.com/s2wz9x for more.

Lab death

A 25-year-old lab worker who was studying *Neisseria meningitidis* died on 28 April, apparently from an infection with the bacterium. Richard Din was researching vaccines against subtype B of *N. meningitidis* (a strain for which there is no effective vaccine) at the Veterans Affairs Medical Center in San Francisco, California. The news was revealed on 2 May; state and federal health and safety officials are investigating Din's death. See go.nature.com/w2v6uc for more.

Science chief leaves

Bruce Alberts will depart his post as editor-in-chief of the journal *Science* by March 2013, said its publisher, the American Association for the Advancement of Science in Washington DC, on 3 May. Alberts (**pictured**) is professor

T. KOCHEL/AAAS

emeritus of biochemistry and biophysics at the University of California, San Francisco, and became the 18th editor-in-chief of the journal in 2008. See go.nature.com/r7eiqi for more.

Retiring president

The US National Academy of Engineering (NAE) president Charles Vest, a mechanical engineer who took office in 2007, will step down next year and not serve a second six-year term, the academy confirmed last week. The organization is looking for a successor. Before his stint at the NAE, Vest was president of the Massachusetts Institute of Technology in Cambridge for 14 years.

German plagiarism

German research minister Annette Schavan is the latest target in the country's recent rash of plagiarism accusations against high-level politicians. The University of Düsseldorf is investigating accusations posted on an anonymous webpage on 2 May that Schavan had inadequately cited original sources in her 1980 thesis on moral education. Some experts say that the extent of the alleged plagiarism is minor and have called for a formal, independent body to assess future allegations to avoid witch-hunts.

New to US academy

Among the 84 members elected to the US National Academy of Sciences on 2 May were Karl Deisseroth, an optogenetics researcher at Stanford University in California, and planetary scientist Robin Canup of the Southwest Research Institute in Boulder, Colorado. In total, the academy elected 26 women as members — the most it has ever admitted in one year. The academy also added 21 foreign associates; it now has 2,152 active members and 430 foreign associates. See go.nature.com/onztjx for more.

RESEARCH

Jupiter mission

The European Space Agency (ESA) has approved a roughly €1-billion (US$1.3-billion) mission to Jupiter and its moons, scheduled to launch in 2022. On 2 May, ESA's Science Programme Committee endorsed the Jupiter Icy moons Explorer, or JUICE, as its next 'large class' space probe. The 11-year mission would explore Jupiter's aurora and hunt for liquid oceans beneath the surfaces of the moons Ganymede and Europa. JUICE was picked over two other missions — the space-based New Gravitational wave Observatory, and an X-ray telescope called ATHENA. See go.nature.com/lrvyl2 and page 148 for more.

Cancer-genome hub

The University of California, Santa Cruz, has opened a US$10.3-million cancer-genomics repository for data from the three major US cancer-sequencing projects: the Cancer Genome Atlas, the Therapeutically Applicable Research to Generate Effective Treatments project and the Cancer Genome Characterization Initiative. That makes the database the largest collection of cancer genomes accessible to researchers around the world. Aggregating the data into one place is key to advancing the use of personal genomics technologies in cancer, says project leader David Haussler. See go.nature.com/xhnj9l for more.

COMING UP

14–15 MAY
The US National Institutes of Health hosts a summit to discuss the latest research on Alzheimer's disease, in Bethesda, Maryland. go.nature.com/eo9nyp

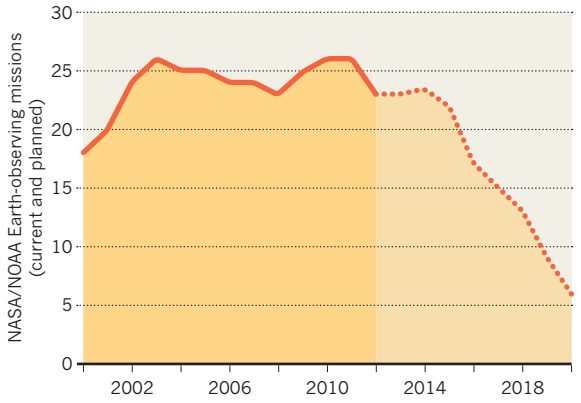
14–15 MAY
In Washington DC, the US National Science Foundation hosts a global summit to discuss principles and procedures for peer review.

TREND WATCH

Five years ago, a decadal survey of the United States' space-based environmental-monitoring programme by the National Research Council (NRC) found that its satellite system was "at risk of collapse". In a 2 May study, the NRC says that the outlook is worse: rising costs and the launch failures of NASA's Orbiting Carbon Observatory (February 2009) and Glory satellite (March 2011) mean the US Earth-observing system is beginning a "rapid decline in capability". See go.nature.com/p6xxhv for more.

UNITED STATES TO LOSE EARTH-OBSERVING POWER

Delays, mission changes and budgetary woes mean a steep decline for US satellite missions to observe Earth.



SOURCE: NRC

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

PHYSICS Will a fog of data make the Higgs boson harder to see? **p.156**

PHYSICS Claim about quantum waves puts theorists in a spin **p.157**

BIOTECHNOLOGY First drug made in plant cells wins approval **p.160**

EPIDEMIOLOGY Bid to rescue archives of radiation data gathers steam **p.162**

MARTIN SCHOELLER/AUGUST

Culture shock: Pirahã speakers may be deviating from the largely accepted theory of language.

LINGUISTICS

War of words over tribal tongue

Debate highlights pitfalls in studying minority languages.

BY EUGENIE SAMUEL REICH

It wasn't long after his translation of the Gospel of St Mark failed to interest the Pirahã tribe members he was trying to convert to Christianity that Daniel Everett, then a missionary and linguistic anthropologist, began to doubt what he had learned about the foundations of human language.

Thirty years on, Everett, now at Bentley University in Waltham, Massachusetts, has long since left missionary work, but his study of the Pirahã tongue has increasingly cast him in the role of heretic in a battle over the influence of culture in shaping the structure of a language. The debate has resurfaced with the publication in March of his book *Language: The Cultural Tool* and a related television documentary scheduled to be broadcast this week in the United States. But as Everett's controversial views gain attention, other scholars are beginning to question his interpretations.

When Everett began to learn Pirahã — today spoken by fewer than 400 people in the interior Brazilian state of Amazonas — he expected it to share certain grammatical features with other languages. These features, he says, would make Pirahã consistent with the concept of a 'universal grammar', which Noam Chomsky, a linguist at the Massachusetts Institute of Technology (MIT) in Cambridge, has famously theorized is hard-wired into the human brain. Over time, however, Everett concluded that Pirahã was missing some of those supposedly universal features, including the use of embedded clauses. In most languages, such clauses serve a wide range of functions, allowing speakers to discuss the thoughts of others, for example.

Everett also says that Pirahã speakers are reluctant to generalize beyond direct experience, or to talk about people they do not know, perhaps explaining their lack of interest in the biblical figures of his translation. He eventually concluded that these differences arose from the Pirahã having a culture that it based in the 'here and now', and he argues that this culturally determined grammar conflicts directly with Chomsky's theory of language.

Because Everett has spent far more time than anyone else living among the Pirahã and studying their language (some eight years, by his estimate), it has been difficult for other researchers to evaluate his claims, says Jan-Wouter Zwart, a linguist at the University of Groningen in the Netherlands. "All I know about Pirahã is from his grammar, and that's true for all of us. We are typically dependent on a single person's work."

Now, however, another researcher has collected independent data on Pirahã, and he says that his findings do not support Everett's interpretation. At a presentation in April at MIT, Uli Sauerland, a linguist at the Centre for General Linguistics in Berlin, told the audience: "My evidence is that they can express attitudes, and what I think they use to do this is embedded sentences." Sauerland is now preparing his data for publication.

In one experiment, Sauerland showed Pirahã speakers a skit in which one actor moved an object — such as a papaya or nut — from one hiding place to another, in front of a second, blindfolded actor. Sauerland says that the remarks of the subjects, when asked to describe the skit, could be best translated as: "Oope thinks the nut is under the banana leaf. It is really under the basket", or "Oope doesn't know where the nut is", in which the parts of the sentence describing Oope's thoughts are embedded clauses. ▶

NATURE.COM
To hear a native Pirahã speaker, visit:
go.nature.com/7lrcoy

Although Everett didn't see Sauerland's presentation, he suggests that the remarks could equally well be translated as two separate, direct sentences, such as: "The nut is under the banana leaf. Or so Oope says", or "Oope doesn't know. Where is the nut?"

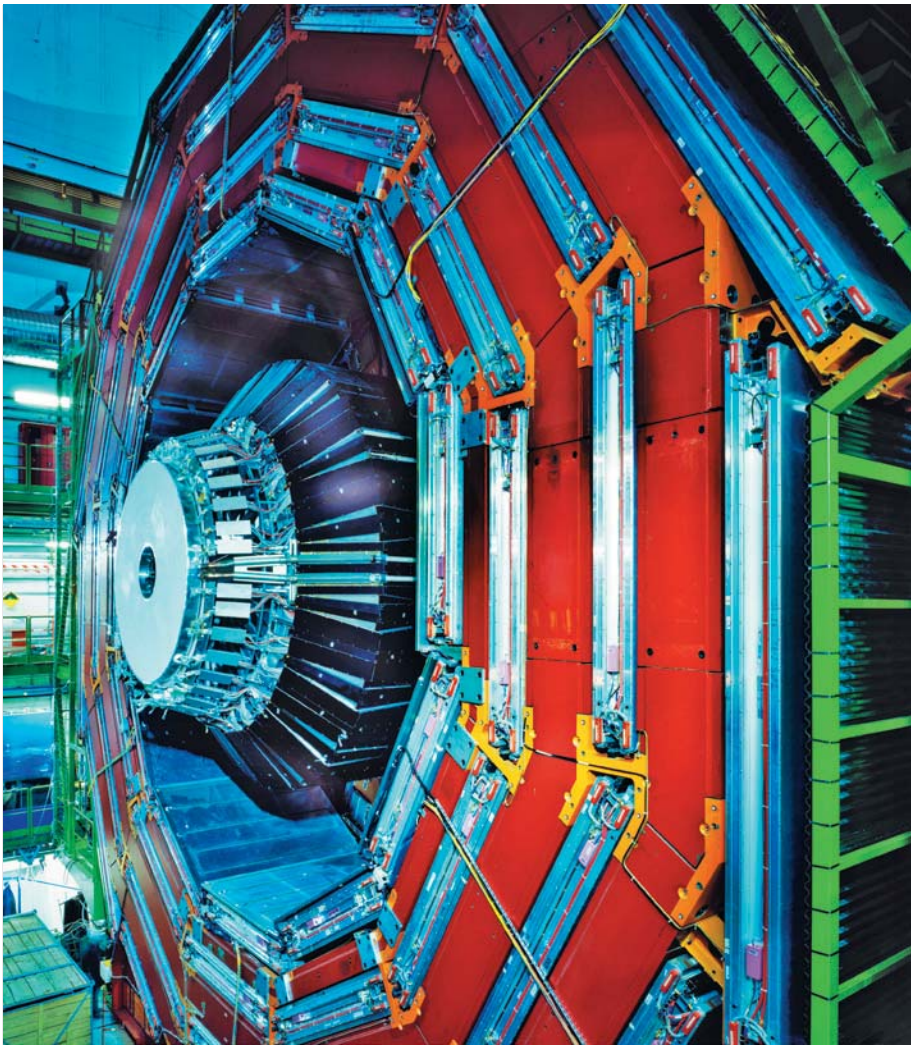
The difference is subtle, but it cuts to the heart of Everett's case against Chomsky's theory. Embedded clauses can be instances of recursion, an iterative process that Chomsky says is essential to all language because it enables ever more complex sentences to be built up out of individual words or sounds. Everett also says that Pirahã lacks colour and number terms and has no perfect tense, which is used in English for events that have been completed. Although many linguists say that Chomsky's theory of a universal grammar would hold even if Everett is right about those features, Everett believes that such a profound interplay between culture and language conflicts with Chomsky's theory of language as innate.

The situation underscores the potential difficulty in settling important claims about minority languages. The United Nations Educational, Scientific and Cultural Organization lists 2,473 languages as endangered, meaning either that they are spoken by only small communities of people or that the elderly people who speak them have not passed them on to subsequent generations. Many such languages have been studied by just a single linguist, so that other researchers must rely on that person's translations.

"For a lot of languages we have extremely poor documentation," says Lyle Campbell, a linguist at the University of Hawaii at Manoa who is leading ELCat, an online project supported by the US National Science Foundation that aims to catalogue endangered languages. Expected to launch later this month, ELCat will serve as a centralized repository for original data such as recordings, video, text, transcripts and translations. Campbell says that such documentation makes it possible for linguists to test each others' statements.

Thomas Roeper, a linguist at the University of Massachusetts in Amherst, says that linguists will inevitably have to work with data from a limited number of sources. "There are many languages that only one, two or three people have studied, with Western prejudices. It would be a great mistake if we didn't include their experiences in our knowledge," he says.

Everett and his colleagues are now testing his arguments using data on Pirahã collected by his missionary predecessor, Steve Sheldon. Everett is also working on making his own records available. "I have data recorded, and am translating more and more," he says. ■



The Compact Muon Solenoid experiment detects hundreds of millions of particle collisions every second.

ENRICO SACCHETTI

PARTICLE PHYSICS

LHC prepares for data pile-up

Physicists scramble to see through fog of collisions.

BY GEOFF BRUMFIEL

The world's largest particle accelerator is roaring along at an unprecedented pace, delivering torrents of data to its physicist handlers. But the hundreds of millions of collisions happening inside the machine every second are now growing into a thick fog that, paradoxically, threatens to obscure a fabled quarry: the Higgs boson.

The problem is known as pile-up, and it promises to be one of the greatest challenges this year for scientists working on the Large Hadron Collider (LHC) at CERN, Europe's main high-energy physics laboratory near Geneva, Switzerland.

Huge amounts of computing power, cunning software and technical tricks are helping scientists to stay ahead of the problem. But researchers may still need to scale back the collisions to find the long-sought Higgs, the manifestation of a field that is believed to confer mass on other particles.

If it exists, the Higgs will appear fleetingly inside the machine before decaying into lighter particles. Last year, the two biggest detectors at the LHC saw hints of a Higgs with a mass of about 125 gigaelectronvolts (energy and mass are interchangeable in particle physics). This year, researchers want to collect more data to see whether that signal grows into a certainty, or withers back to nothing.

Since it began its latest science run last month, the LHC has been squeezing trillions of protons into ever-smaller bunches, and smashing those bunches

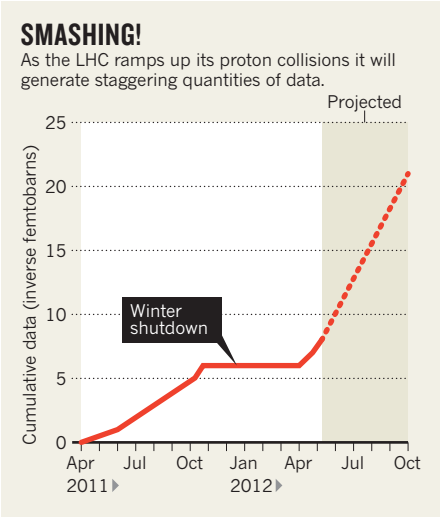
NATURE.COM
To read more articles on the LHC, see:
go.nature.com/uzvtfd

together tens of millions of times per second. The resultant data are measured in inverse femtobarns (fb^{-1}), a unit roughly equivalent to 100 trillion collisions. In the past month alone, the LHC recorded 1 fb^{-1} worth of collisions. By the end of the year it aims have captured at least 15 fb^{-1} (see 'Smashing!').

To gather these data, researchers are pushing the collider in two ways: by accelerating the particles to ever-greater energies and by increasing the number of collisions. Higher energies allow heavier particles to pop into being, but it is the number of collisions that will determine whether physicists have enough data to declare a discovery. In the weeks ahead, scientists will pack more protons inside the machine and focus the particles as tightly as possible onto the collision points at the centre of the LHC's two biggest detectors. Already, "we've done humongously better than we thought we could", says Mike Lamont, the head of accelerator operations at CERN.

Every time two tightly packed bunches of protons cross, they generate not one collision, but on average 27, Lamont says. But within a few weeks, that number is expected to rise into the mid-30s, peaking at around 40 collisions per crossing. The two main detectors at the LHC were designed to handle only around two dozen collisions at once. But they have managed to cope so far.

Each detector is made up of layers of smaller detectors that record the tracks of debris coming from their centre. When a collision occurs, computers above the machine decide whether the data are interesting and, if so, reconstruct the collision from the tracks. But when dozens of collisions occur at once, the computers must disentangle them.



SOURCE: CERN

Last year, researchers working with the ATLAS detector formed a task force to tackle the pile-up problem, rewriting computer code so that the detector could cope with the extra collisions. Team member Andreas Salzburger says that the group has been working hard to weed out the 'ghost' particles that appear when the paths of several particles align, creating the illusion of a particle that is not actually there. Eliminating these ghosts as early as possible reduces the amount of computing power needed to crunch useful data, he says.

At the Compact Muon Solenoid (CMS), ATLAS's rival detector, physicists have trained their algorithms to triage data on the fly, analysing particle tracks in order of complexity. "Did you ever play the game 'pick-up sticks?'," asks Joe Incandela, the spokesman for the CMS. "You pick up the easiest ones first, and it makes it simpler to deal with the other ones." The team is also working on ways to get rid of signals from 'loopers', low-energy particles that spiral along the detector's magnetic field lines, generating data that are irrelevant to the Higgs hunt.

Such tricks are likely to be less effective as the number of collisions rise. At the outer edges of the machine, the detector segments are larger and have coarser resolution, so it might not be possible to disentangle some of the tracks. That could reduce a detector's ability to pick up one signature of the Higgs: a decay to a pair of W bosons, which causes a cascade of particles that need to be caught by these outer segments.

For now, the mountains of extra data should offset what is lost to pile-up. Researchers expect to miss no more than 15% of events from the most likely Higgs decay pathway, which produces two γ-rays. And if ATLAS and the CMS can't handle the extra particles surging through the machine, Lamont says, the accelerator physicists are ready to dial it back. But "if they can take it, we will give it to them", he says. ■

PHYSICS

A boost for quantum reality

Theorists claim they can prove that wavefunctions are real states.

BY EUGENIE SAMUEL REICH

The philosophical status of the wavefunction — the entity that determines the probability of different outcomes of measurements on quantum-mechanical particles — would seem to be an unlikely subject for emotional debate. Yet online discussion of a paper claiming to show mathematically that the wavefunction is real has ranged from ardently star-struck to downright vitriolic since the article was first released as a preprint in November 2011.

The paper, thought by some to be one of the most important in quantum foundations in decades, was finally published last week in *Nature Physics* (M. F. Pusey, J. Barrett & T. Rudolph *Nature Phys.* http://dx.doi.org/10.1038/nphys2309; 2012), enabling the authors, who had been concerned about violating the journal's embargo, to speak about it publicly for the first time. They say that the mathematics leaves no doubt that the wavefunction is not just a statistical tool, but rather, a real, objective state of a quantum system. "People have become emotionally attached to positions that they defend with vague arguments," says Jonathan Barrett, one of the authors and a physicist at Royal Holloway, University of London. "It's better to have a theorem."

The authors have some heavyweights in their corner: their view was once shared by Austrian physicist and quantum-mechanics pioneer Erwin Schrödinger, who proposed in his famous thought experiment that a quantum-mechanical cat could be dead and alive at the same time. But other physicists have favoured an opposing view, one held by Albert Einstein: that the wavefunction reflects the partial knowledge an experimenter has about a system. In this interpretation, the cat is either dead or alive, but the experimenter does not know which. This 'epistemic' interpretation, many physicists and philosophers argue, better explains the phenomenon of wavefunction collapse, in which a quantum state is fundamentally changed by measuring it.

Barrett and his colleagues are following the approach of physicist John Bell, who in 1964 proved that quantum mechanics has another counterintuitive implication: that measurements on one particle can influence the state of another, distant particle, faster than the speed of light should allow. Bell's was a 'no-go' theorem: its strategy was to show that theories that do not allow faster-than-light ▶

influences cannot reproduce the predictions of quantum mechanics. Similarly, the theorem proposed by Barrett and his colleagues shows that theories that treat the wavefunction in terms of lack of knowledge of a system's physical state will also fail to reproduce those predictions. Given how well-confirmed quantum mechanics is, the theorem suggests that such epistemic theories are wrong. "I hope this will take its place alongside Bell's theorem," says Barrett.

GROUNDED IN REALITY

If the wavefunction simply reflects the experimenter's uncertainty, then different wavefunctions could represent the same underlying reality, says Terry Rudolph, an author on the paper and a physicist at Imperial College London. Rudolph gives the example of a die that can be prepared to give either even numbers, with a 1/3 probability of getting 2, 4 or 6; or prime numbers, with a 1/3 probability of getting 2, 3 or 5. The real state 2 can be produced by either preparation method, so the same reality underlies two different probabilistic models. The authors show, however, that the same reality cannot underpin different quantum states.

Their theorem does, however, depend on a controversial assumption: that quantum systems have an objective underlying physical state. Christopher Fuchs, a physicist at the Perimeter Institute in Waterloo, Canada, who has been working to develop an epistemic interpretation of quantum mechanics, says that he has avoided the interpretations that the authors exclude. The wavefunction may represent the experimenter's ignorance about measurement outcomes, rather than the underlying physical reality, he says. The new theorem doesn't rule that out.

Still, Matt Leifer, a physicist at University College London who works on quantum information, says that the theorem tackles a big question in a simple and clean way. He also says that it could end up being as useful as Bell's theorem, which turned out to have applications in quantum information theory and cryptography. "Nobody has thought if it has a practical use, but I wouldn't be surprised if it did," he says.

Because it is incompatible with quantum mechanics, the theorem also raises a deeper question: could quantum mechanics be wrong? Everyone assumes that it reigns supreme, but there is always a possibility that it could be overturned. So Barrett is now working with experimentalists to check predictions that differ between the theory and the epistemic accounts it conflicts with. "We don't expect quantum mechanics would fail this test, but we should still do it," he says. ■

N. GILBERT

Children in Mwandama, Malawi, now have a better chance of living to the age of five.

GLOBAL HEALTH

Development project touts health victory

But critics question data and cost estimates from the Millennium Villages Project.

BY NATASHA GILBERT

For villagers in Mwandama, Malawi, visiting a health worker used to mean a daunting 40-kilometre round trip on foot. So the medical centre that was built in the area as part of the Millennium Villages Project (MVP) last year has improved their quality of life — and their health. Research published this week suggests that the MVP has significantly reduced infant mortality at sites across Africa.

But some researchers have questioned the methods used to quantify the benefits of the project, and demanded that the MVP release its underlying data. "The core of the problem is lack of transparency and careful, independent analysis," says Michael Clemens, a migration and development researcher at the Center for Global Development, an independent research institution in Washington DC.

NATURE.COM
For more on science and development in Africa, see:
nature.com/africa

The MVP, which is spearheaded by Jeffrey Sachs, an economist at Columbia University in New York, aims to lift some of the poorest people in Africa out of poverty and improve their standard of living by boosting health and food security. It intends to help villages at 14 sites across Africa to reach the United Nations' eight Millennium Development Goals (MDGs) by 2015.

Sachs says that many aid projects see limited success because they focus on one area at a time. By contrast, the MVP tackles all the root causes of poverty at once. For example, it simultaneously provides free fertilizer and seeds, builds schools and gives business training to farmers. Funded by cash and in-kind contributions from governments, industry and aid donors, the project is growing in influence. The government of Cameroon is planning to start a similar scheme, for example, using funding from Japan and the UN to boost economic and employment opportunities for 50,000 villagers.

Research published in *The Lancet* (P. Pronyk

et al. Lancet http://dx.doi.org/10.1016/s0140-6736(12)60207-4; 2012) now offers quantitative evidence of the success of the MVP model at nine Millennium Village sites in sub-Saharan Africa, including Mwandama (see 'Health targets'). Between 2006 and 2009, mortality in under-fives fell by an average of 22%, reaching a level roughly two-thirds of that in control villages not involved with the project, where child mortality seemed to rise.

In rural areas nationwide, under-five death rates fell by an average of 2.6% each year over the course of a decade — a stark contrast with the Millennium Villages' average of 7.8% for each year of the study. "The consistency with which child deaths went down and the size of the drop was surprising," says Paul Pronyk, director of monitoring and evaluation for the MVP at Columbia, and lead author of the paper.

But Clemens says that these headline figures are misleading for a number of reasons. He points out that the control-village data include retrospectively estimated figures that are probably too high. And nationwide improvements in child mortality over the three years of the study were almost as good as in the Millennium Villages, he says, so it is unfair to compare the project's success with a more gradual decadal trend. Furthermore, deriving trends from children monitored in a few villages for just three years introduces significant statistical uncertainty, he argues.

Using figures in the paper, Clemens calculates that the study authors can be confident only that the annual rate of decline for child mortality in the Millennium Villages lies between 1.4% and 14.3%. "If you claim to triple rates of decline you must have the evidence to back this up," he says.

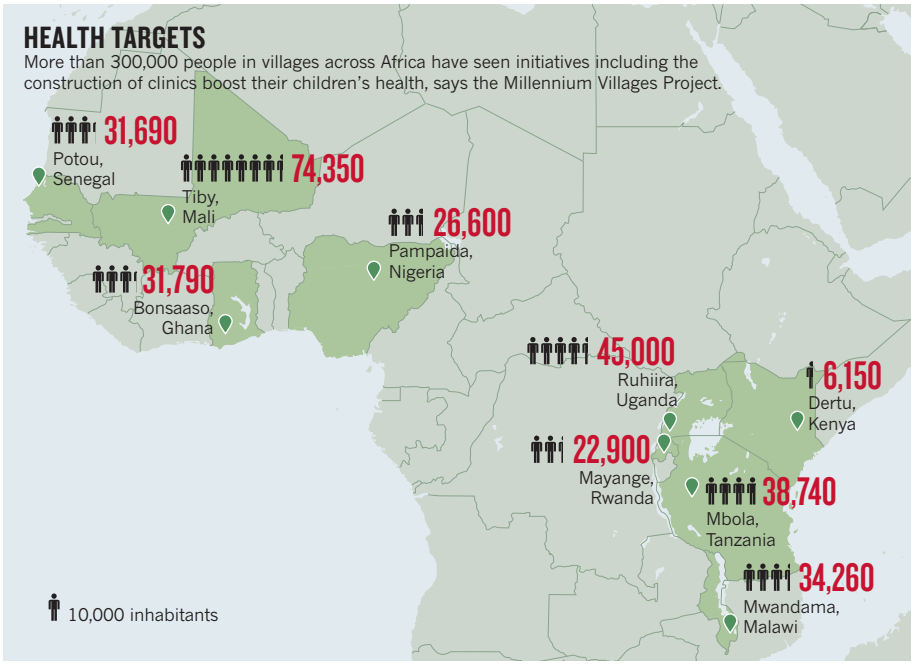
Clemens calls on the MVP to make its raw survey data available for independent scrutiny. He also questions the MVP's stated cost of US$120 per person per year, saying that it is an underestimate because it neglects indirect costs, such as the services of nearby non-governmental organizations.

Pronyk says that the mortality rates are meant to be "illustrative" of general improvement in child health, and he stands by the estimated budget. The MVP makes raw data available when requested by journals, and shares them with external partners, he adds. A Millennium Village scheme starting up in Ghana will be evaluated by independent researchers on behalf of the UK government's Department for International Development, which may help to settle the debate.

Some development experts query how widely the Millennium Village concept can be applied. Stephen Carr, an agriculture and development consultant in Zomba, Malawi, says that schools in the Millennium Villages currently attract the best teachers in the country, boosting children's educational attainment. If such projects become more widespread, they will not have the resources to reach such great heights, he warns.

Pedro Sanchez, an agricultural scientist and co-director of the MVP at Columbia, argues that "scaling up is the business of governments". His project's goal is to establish proof of concept and demonstrate that villages can achieve the MDGs with appropriate resources.

Boosting opportunities in even one school is better than in none at all, adds Andrew Daudi, team leader and science coordinator for the MVP in Mwandama, who is optimistic about the children being educated at the new primary school there. "Some of the children here are the best in Malawi," he says. "The children we educate today will be able to take over the country later." ■ SEE EDITORIAL P.147



SOURCE: P. PRONYK *ET AL.*

BIOTECHNOLOGY

Drug-making plant blooms

Approval of a 'biologic' manufactured in plant cells may pave the way for similar products.

BY AMY MAXMEN

It was midnight when an anxious Ari Zimran finally got the phone call for which he had been waiting. The news couldn't have been better: the drug he had worked on for nearly a decade had just been approved by the US Food and Drug Administration (FDA).

Zimran, who heads the Gaucher Clinic in Jerusalem and is a member of the scientific advisory board at Protalix Biotherapeutics, a small biotechnology firm in Carmiel, Israel, was not the only one celebrating the company's success last week. Biotechnologists around the world cheered, because Protalix's Elelyso (taliglucerase alfa) is the first biological drug for human use that is manufactured inside modified plant cells.

"It's a great day for plant-made pharmaceuticals," says Scott Deeter, president of Ventria Bioscience, a biotech firm based in Fort Collins, Colorado. "This shows the triumph of innovators over the status quo, and that's really very important."

Drugs that are based on large biological molecules — known as biologics — have been produced inside genetically engineered animal cells, yeast and bacteria for more than two decades. Insulin has been made by genetically modified *Escherichia coli* bacteria since 1982, and by 2010, the global market for such therapies had reached about US$149 billion.

Since the early 1990s, some researchers have been developing plants that could act as cheaper factories for biologics. Plant-cell cultures are also attractive because they require less precise conditions for growth than animal cells. But efforts to exploit plants in this way have lagged, in part because companies and investors were wary of this unfamiliar production method.

Protalix was strategic in targeting a rare heritable disorder called Gaucher's disease, because current means of producing treatments for it have fallen short. The disease is caused by an enzyme malfunction that results in the accumulation of fat in cells and organs, with symptoms ranging from bone deterioration to anaemia. Two existing drugs compensate for the enzyme deficiency, but they can cost up to $300,000 per year in the United States, and drug shortages in recent years have left some patients in need of hospital care.

PLANTS IN THE PIPELINE
Manufacturers have begun or completed phase II clinical trials on a handful of biologics made in plants, and hope to follow Elelyso to market.

Drug	Condition	Company	Platform
Locteron (interferon-α)	Hepatitis C	Biolex Therapeutics	Duckweed
H5N1 vaccine	Influenza	Medicago	Tobacco
VEN100	Antibiotic-associated diarrhoea	Ventria Bioscience	Rice
CaroRx	Dental caries	Planet Biotechnology	Tobacco

Structurally, Protalix's Elelyso resembles one of those drugs: Cerezyme, made by Genzyme in Cambridge, Massachusetts. Cerezyme is produced in modified hamster cells, which require regulated temperatures, a complex growth solution and an environment scrubbed free of the viruses that infect hamsters and humans alike. These factors contributed to manufacturing problems that dogged Genzyme last year, limiting supplies of Cerezyme.

"It's really the regulatory hurdles and costly clinical trials that are a barrier."

Protalix's solution is to take a normal version of the human gene affected in Gaucher's disease and introduce it into carrot cells, which are more robust than hamster cells, and then extract the enzyme they make. The lower production overheads will allow the company to sell Elelyso for just 75% of the price of Cerezyme, the most popular drug on the market, says David Aviezer, Protalix's president.

Charles Arntzen, a plant biotechnologist at Arizona State University in Tempe, says that Elelyso's approval sends a clear and positive signal to investors and companies that plant-manufactured drugs are worth pursuing (see 'Plants in the pipeline'). When he began working on plant-made vaccines in 1991, he says that he was naive about how long it would take for the technology to blossom. He expected companies and the FDA to embrace the technique, speeding inexpensive products to market.

"Many of us in academia thought that manufacturing costs were a significant part of the entry barrier in making a new product," Arntzen says. But "it's really the regulatory hurdles and costly clinical trials that are a barrier, and big pharmaceutical companies don't want to take this on because they know there is an enormous risk inherent to trying something new".

For those companies trying to produce drugs from whole plants, rather than in cultures of plant cells, Aviezer cautions that Elelyso's approval might not set a precedent. But others in the field are more optimistic. "Even though [Protalix's] technology doesn't use whole plants, it does address many issues of producing proteins in plant cells," says molecular immunologist Julian Ma of St George's, University of London, who is scientific coordinator for Pharma-Planta, a European consortium that is developing plant-derived pharmaceuticals to treat, for example, HIV (see *Nature* **458,** 951; 2009).

Nathalie Charland of Canadian biotech company Medicago, in Quebec City, which is developing vaccines produced in tobacco plants, agrees: "I don't think there will be major differences in how the FDA handles their product and ours." ■

CORRECTION
The News Feature 'Date with history' (*Nature* **485,** 27–29; 2012) incorrectly located the University of Waikato in Wellington instead of Hamilton.

MORE ONLINE

TOP STORY

SEAN MYLES

Solomon Islanders evolved blonde hair separately go.nature.com/uatj9k

MORE NEWS

- Remote-controlled genes trigger insulin production go.nature.com/rtlsod
- Disputed ancient bones in California stay put for now go.nature.com/i2foes
- Call for standards in egg bio-monitoring go.nature.com/qc6zs8

BIOTECHNOLOGY

Drug-making plant blooms

Approval of a 'biologic' manufactured in plant cells may pave the way for similar products.

BY AMY MAXMEN

It was midnight when an anxious Ari Zimran finally got the phone call for which he had been waiting. The news couldn't have been better: the drug he had worked on for nearly a decade had just been approved by the US Food and Drug Administration (FDA).

Zimran, who heads the Gaucher Clinic in Jerusalem and is a member of the scientific advisory board at Protalix Biotherapeutics, a small biotechnology firm in Carmiel, Israel, was not the only one celebrating the company's success last week. Biotechnologists around the world cheered, because Protalix's Elelyso (taliglucerase alfa) is the first biological drug for human use that is manufactured inside modified plant cells.

"It's a great day for plant-made pharmaceuticals," says Scott Deeter, president of Ventria Bioscience, a biotech firm based in Fort Collins, Colorado. "This shows the triumph of innovators over the status quo, and that's really very important."

Drugs that are based on large biological molecules — known as biologics — have been produced inside genetically engineered animal cells, yeast and bacteria for more than two decades. Insulin has been made by genetically modified *Escherichia coli* bacteria since 1982, and by 2010, the global market for such therapies had reached about US$149 billion.

Since the early 1990s, some researchers have been developing plants that could act as cheaper factories for biologics. Plant-cell cultures are also attractive because they require less precise conditions for growth than animal cells. But efforts to exploit plants in this way have lagged, in part because companies and investors were wary of this unfamiliar production method.

Protalix was strategic in targeting a rare heritable disorder called Gaucher's disease, because current means of producing treatments for it have fallen short. The disease is caused by an enzyme malfunction that results in the accumulation of fat in cells and organs, with symptoms ranging from bone deterioration to anaemia. Two existing drugs compensate for the enzyme deficiency, but they can cost up to $300,000 per year in the United States, and drug shortages in recent years have left some patients in need of hospital care.

PLANTS IN THE PIPELINE
Manufacturers have begun or completed phase II clinical trials on a handful of biologics made in plants, and hope to follow Elelyso to market.

Drug	Condition	Company	Platform
Locteron (interferon-α)	Hepatitis C	Biolex Therapeutics	Duckweed
H5N1 vaccine	Influenza	Medicago	Tobacco
VEN100	Antibiotic-associated diarrhoea	Ventria Bioscience	Rice
CaroRx	Dental caries	Planet Biotechnology	Tobacco

Structurally, Protalix's Elelyso resembles one of those drugs: Cerezyme, made by Genzyme in Cambridge, Massachusetts. Cerezyme is produced in modified hamster cells, which require regulated temperatures, a complex growth solution and an environment scrubbed free of the viruses that infect hamsters and humans alike. These factors contributed to manufacturing problems that dogged Genzyme last year, limiting supplies of Cerezyme.

"It's really the regulatory hurdles and costly clinical trials that are a barrier."

Protalix's solution is to take a normal version of the human gene affected in Gaucher's disease and introduce it into carrot cells, which are more robust than hamster cells, and then extract the enzyme they make. The lower production overheads will allow the company to sell Elelyso for just 75% of the price of Cerezyme, the most popular drug on the market, says David Aviezer, Protalix's president.

Charles Arntzen, a plant biotechnologist at Arizona State University in Tempe, says that Elelyso's approval sends a clear and positive signal to investors and companies that plant-manufactured drugs are worth pursuing (see 'Plants in the pipeline'). When he began working on plant-made vaccines in 1991, he says that he was naive about how long it would take for the technology to blossom. He expected companies and the FDA to embrace the technique, speeding inexpensive products to market.

"Many of us in academia thought that manufacturing costs were a significant part of the entry barrier in making a new product," Arntzen says. But "it's really the regulatory hurdles and costly clinical trials that are a barrier, and big pharmaceutical companies don't want to take this on because they know there is an enormous risk inherent to trying something new".

For those companies trying to produce drugs from whole plants, rather than in cultures of plant cells, Aviezer cautions that Elelyso's approval might not set a precedent. But others in the field are more optimistic. "Even though [Protalix's] technology doesn't use whole plants, it does address many issues of producing proteins in plant cells," says molecular immunologist Julian Ma of St George's, University of London, who is scientific coordinator for Pharma-Planta, a European consortium that is developing plant-derived pharmaceuticals to treat, for example, HIV (see *Nature* **458,** 951; 2009).

Nathalie Charland of Canadian biotech company Medicago, in Quebec City, which is developing vaccines produced in tobacco plants, agrees: "I don't think there will be major differences in how the FDA handles their product and ours." ■

CORRECTION
The News Feature 'Date with history' (*Nature* **485,** 27–29; 2012) incorrectly located the University of Waikato in Wellington instead of Hamilton.

MORE ONLINE

TOP STORY

SEAN MYLES

Solomon Islanders evolved blonde hair separately go.nature.com/uatj9k

MORE NEWS

- Remote-controlled genes trigger insulin production go.nature.com/rtlsod
- Disputed ancient bones in California stay put for now go.nature.com/i2foes
- Call for standards in egg bio-monitoring go.nature.com/qc6zs8

Signs near the spent-fuel processing facility at Russia's Mayak Nuclear Power Plant forbid entrance to the contaminated area.

I. YAKOVLEV/ITAR-TASS

RAIDERS OF THE LOST ARCHIVE

Old collections of irradiated tissues could answer modern-day questions about the dangers of radiation. Now, researchers are making a concerted effort to save the stores.

BY ALISON ABBOTT

The town of Ozersk, deep in Russia's remote southern Urals, hides the relics of a massive secret experiment. From the early 1950s to the end of the cold war, nearly 250,000 animals were systematically irradiated. Some were blasted with α-, β- or γ-radiation. Others were fed radioactive particles. Some of the doses were high enough to kill the animals outright; others were so low that they seemed harmless. After the animals — mice, rats, dogs, pigs and a few monkeys — died, scientists dissected out their tissues to see what damage the radioactivity had wrought. They fixed thin slices of lung, heart, liver, brain and other organs in paraffin blocks, to be sliced and examined under the microscope. Some organs, they pickled in jars of formalin.

Fearful of a nuclear attack by the United States, the Soviet Union wanted to understand how radiation damages tissues and causes diseases such as cancer. Concerns about home-grown accidents, such as the 1957 disaster at the Mayak nuclear plant close to Ozersk, were another motivation. Throughout their experiments, the scientists carefully preserved the tissues and meticulously recorded their findings. Similar archives of irradiated tissue were built up in the United States, Europe and Japan, where nearly half a billion animals were sacrificed to the cause. But when the cold war came to an end, the collections fell into disrepair.

Now, these archives have become important to a new generation of radiobiologists, who want to explore the effects of the extremely low doses of radiation — below 100 millisieverts — that people receive during medical procedures such as computed-tomography diagnostic scans, and by living close to the damaged Fukushima nuclear reactors in Japan.

The old collections provide a resource that could not be recreated today. Most of the experiments were done under precise conditions, at a wide range of radiation doses and usually for the lifetime of the animals. "We will never be able to repeat the scale of those animal experiments, for both funding and ethical reasons," says Gayle Woloschak, a radiation biologist at Northwestern University in Chicago, Illinois. "But maybe we can reuse the legacy tissue." Over the past few years, researchers around the world have organized an effort to identify and save tissue archives from all the major animal irradiation experiments, and they have won support from a diverse range of funding agencies, including the European Commission, the US National Cancer Institute and the US Department of Energy.

But the challenges are still great. Researchers have to show that the age of the samples, and the preservation techniques used on them, have not affected the DNA, RNA and proteins the samples contain. They have to piece together such molecular data to reveal whether cell circuitry is disrupted at low radiation doses. Their early tests are indicating that some of the samples will be usable, making them regret how much of such painstakingly collected material around the world has already been lost.

RADIATION RESERVOIRS

When the ageing survivors of the Hiroshima and Nagasaki nuclear bombs and the contaminated Mayak workers started to develop cardiovascular disease at above-normal rates[1,2], it became clear that radiation does more than just cause cancer. What is not known is whether or how very low doses of radiation might increase the risk of these and other diseases. Biologists have generally assumed that the damage will be proportional to the dose, but *in vitro* studies have shown that cells can

repair modest DNA damage caused by radiation — and that low-dose radiation might even protect the cell against future exposure.

"Maybe there is a threshold dose below which radiation is not harmful," says Wolfgang Weiss, head of radiation protection and health at Germany's Federal Office for Radiation Protection in Munich. Epidemiological studies on people exposed to radiation through their jobs, nuclear accidents or medical procedures haven't shed much light on the matter. Some of the studies contained too few people to detect what could be a tiny increased incidence in disease; in others, it is unclear what dose the individuals received. So although radiation protection agencies typically restrict occupational exposure (for the nuclear industry, for example) to an average of 20 mSv per year, scientists don't have hard data on which to base high-stake conclusions about what level of radiation, if any, is really safe. The old animal tissues could hold some of the answers.

In February 2007, the quest to find such tissues took Soile Tapio on a mission from one of Germany's former nuclear research centres, the Helmholtz Centre Munich, to a dark, frigid Ozersk. Tapio was taking part in a programme called the Promotion of the European Radiobiology Archives (ERA-PRO), part of an effort dating back to 1996 to digitize the data from radiation experiments done in Europe. In 2006, the director of the animal irradiation programme at the Southern Urals Biophysics Institute (SUBI) in Ozersk alerted Tapio to the enormous scope of the studies there. "At the time we didn't know much more about SUBI than its name," Tapio says. She certainly hadn't known quite what to expect when she set off there with her small ERA-PRO delegation.

"WE WILL NEVER BE ABLE TO REPEAT THE SCALE OF THOSE ANIMAL EXPERIMENTS."

S. TAPIO

Samples from 23,000 animals irradiated during the cold war are now meticulously archived at the Southern Urals Biophysics Institute.

OFF-LIMITS

It had already taken a few months to get approval from Russia to visit the closed town of Ozersk. After a long flight, a three-hour drive and a lengthy security clearance, a small group of ageing scientists led the delegation to an abandoned house with a gaping roof and broken windows. Glass slides and laboratory notebooks lay strewn on the floors of some offices. But other, heated rooms held wooden cases stacked with slides and wax blocks in plastic bags. In its heyday, the programme had more than 100 staff; but when it was abruptly shut down in the wake of the cold war, just four or five people were left to look after the material it had produced. The visitors were impressed to find that these scientists could link all the samples, from 23,000 animals, to detailed protocols of individual experiments. "The scientists were so happy that at last someone was taking notice of the collection," says Tapio. "They told me many times that they wanted to bring it into order before they died."

Meanwhile, another tissue rescue operation was taking place in the United States. In the mid-1990s, Woloschak had worked on samples from 7,000 beagles and 50,000 mice that had been irradiated in experiments at the Argonne Research Laboratory in Illinois between 1969 and 1992. But after she moved to Northwestern, she was dismayed to hear that the samples were being thrown out and secured permission from the Department of Energy to store them at Northwestern.

"When the community found out I had all the Argonne tissues they began to ask if I could save their tissues, too," Woloschak says. Northwestern University is now the official home for material from all US animal irradiation studies, and Woloschak estimates that she has so far received 20,000 samples. But she has also discovered that many samples have already been destroyed, including those from vast mouse studies done at Oak Ridge National Laboratory in Tennessee and some large-scale dog studies conducted at the University of California, Davis. Woloschak says that she "felt frustrated and angry that the government had invested so many millions of dollars — and immense human effort — into studies that were just going to be trashed because of concerns about space". Tissue collections have also been destroyed elsewhere, including from experiments done at Hiroshima University in Japan, the Italian National Agency for New Technologies, Energy and Sustainable Economic Development's research centre in Casaccia and the UK Medical Research Council's complex in Harwell.

Scientists know that laying their hands on the old tissues will be just the first challenge: they then have to work out whether the biomolecules in the materials can still be detected and measured. They want to identify and analyse the molecular pathways hit by low-dose radiation to see how cells in different tissues adjust — or fail to adjust — to the stress, and how that might set them on the path to disease. They also want to find patterns of biological molecules that might help to determine how much radiation a person received or whether he or she is particularly susceptible to radiation-induced illness.

Woloschak's 1990s work on the old Argonne lab mouse samples provides some hope. She found, for example, that by using a technique called the polymerase chain reaction to amplify genes, she could detect mutations or rearrangements in cancer-specific genes in irradiated tissue that had turned cancerous[3]. Tapio, meanwhile, has adapted standard proteomics techniques so that they can be applied to some of the old tissues, and several groups are studying whether micro-RNAs — which help to control gene expression and are relatively stable — are present in the samples.

Scientists are now poised to apply such work systematically to the legacy tissues. Tapio, for example, is about to start work on paraffin-embedded heart tissue from irradiated mice from the old Russian and US studies. She wants to identify any signs of damage that might explain the elevated incidence of cardiovascular disease in nuclear-bomb survivors. "The scientists who did those studies were only looking for cancer, but we can now look at other diseases we know are relevant," she says.

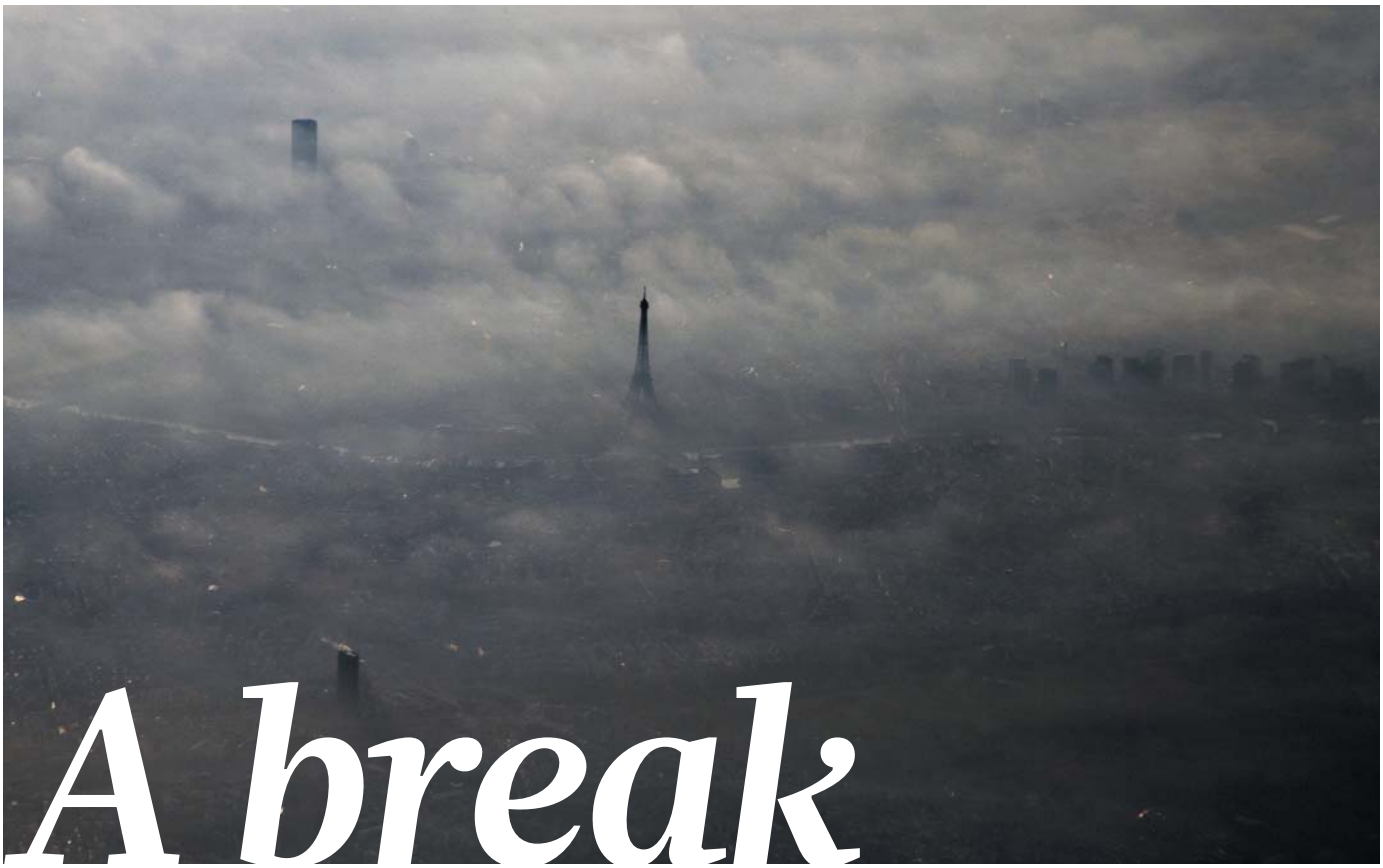
No one is expecting the answers to be quick or simple. The studies could identify many molecular responses that have little to do with disease. "The cell's stress response to any dose of radiation — below that which just fries it — is a complex web of activities, probably affecting many different molecular pathways," says Tapio. And radiobiologists expect that the threshold 'safe' dose will vary between tissues and between individuals.

But at the very least, the tissues in Ozersk have been brought to order, as their guardians hoped. They will soon move into a state-of-the-art storage building being built in the SUBI campus, along with human tissues from radiation-exposed Mayak workers. The animal tissues, researchers hope, will find a new experimental life — this time on an international stage. ■

NATURE.COM
For a podcast on irradiation work, see:
go.nature.com/jeagkg

Alison Abbott *is* Nature*'s senior European correspondent.*

1. Little, M. P. *et al. Radiat. Environ. Biophys.* **49,** 139–153 (2010).
2. Azizova, T. V. *et al. Radiat. Environ. Biophys.* **50,** 539–552 (2011).
3. Haley, B. *et al. Health Phys.* **100,** 613–621 (2011).



H. GRUYAERT/MAGNUM PHOTOS

A break in the clouds

Clouds and aerosol particles have bedevilled climate modellers for decades. Now researchers are starting to gain the upper hand.

BY JEFF TOLLEFSON

Seen from space, Earth can look dressed up or downright dowdy, depending on the location. In some spots, swathes of cloud cloak the dark ocean, offering a stunning contrast of hues. In others, power plants spew out plumes of grey haze and desert storms cover vast regions in palls of dust.

Together, those clouds and the fine particles, which are known as aerosols, do more than just obscure the planet's surface. By reflecting, absorbing and emitting radiation, they have a major role in setting Earth's temperature and have proved maddeningly difficult to simulate in atmospheric models. For decades, they have been the biggest sources of uncertainty in forecasts of future climate.

But researchers say they are beginning to turn a corner in simulating clouds and aerosols. In recent months, climate scientists have started rolling out initial results from the newest generation of models, which represent atmospheric chemistry and microphysics in much more sophisticated ways than previous incarnations. These models allow clouds and aerosols to evolve as they interact with each other and respond to factors such as temperature, relative humidity and air currents. And early results suggest that such processes have a much greater impact on regional climate than scientists had realized. Recent studies have shed light on the roles that clouds and aerosols might have in triggering major African droughts, altering Arctic climate and weakening the monsoon in southern Asia.

"This is fundamentally new science," says Ben Booth, a climate modeller at the UK Met Office Hadley Centre in Exeter, who is investigating how aerosols influence surface temperatures in the North Atlantic Ocean and affect the weather on the surrounding continents. "The new generation of models is changing the kinds of questions we face as scientists."

And more science is coming soon. Leading climate-modelling groups around the world are racing to work up their latest results for the Intergovernmental Panel on Climate Change (IPCC), which is due to release its fifth report section by section in 2013 and 2014. It is already clear that the issue of aerosols and clouds will provide some of the biggest surprises. "This is the real wild card,"

says Ron Stouffer, a climate researcher at the National Oceanic and Atmospheric Administration's Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton, New Jersey.

THE DROUGHT-MAKERS

Each day, the winds that sweep east across North America stir up a witch's brew of atmospheric refuse. Power plants belch out sulphur dioxide gas, which evolves into sulphate particles that reflect sunlight and serve as seeds for clouds. Microscopic specks of carbon rise from vehicles, steel smelters, agricultural fires and other sources. The brighter carbon particles scatter the Sun's rays and dark ones absorb them, processes known as the direct aerosol effect. As the particles ride the air currents eastward, they collide with each other and mix with natural dust and ocean spray to form the load of atmospheric aerosols. Over time, they can build up chemical coatings or merge to form new particles with different properties.

The prevailing winds carry this aerosol stew on a long horseshoe-shaped route around the Atlantic basin (see graphic). The particles are first transported eastward across the ocean, then take a right turn down the coast of France, gathering up more pollution from Europe. The aerosol-laden air curves towards the west coast of North Africa before veering westward and riding tropical air currents back towards America.

Scientists have proposed that this arc of aerosols could block enough sunlight to cool sea surface temperatures in the Atlantic Ocean and alter the regional climate. So Booth and fellow researchers at the Hadley Centre tested the idea with their newest model, which simulates not only the direct aerosol effect but also many of the indirect effects that aerosols have on cloud properties. These interactions take place on too fine a scale to simulate in a global model, so they are represented by statistical equations derived from even more detailed models.

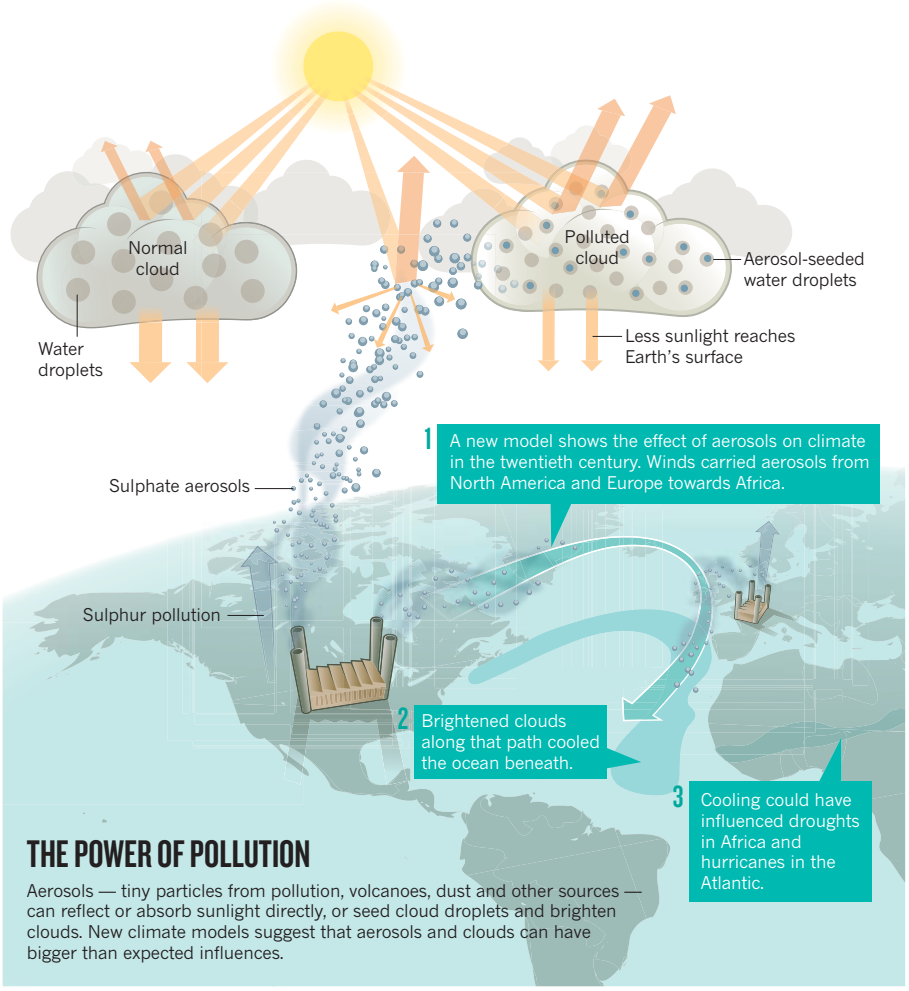
The Hadley Centre team reported last month that, in the model, the aerosols had an exceptionally large effect on North Atlantic sea surface temperatures[1]. And it was an indirect aerosol effect that made the bulk of the difference. The sulphate particles attracted water vapour to create a vast supply of tiny droplets within clouds, brightening them and reducing the amount of sunlight reaching the sea surface.

Overall, North Atlantic sea surface temperatures climbed throughout the simulation, from 1860 to 2005. But an increase in aerosols slowed the ocean warming during the mid-twentieth century, when rapid industrialization caused extreme levels of air pollution. After restrictions on sulphur emissions in the United States and Europe started to kick in the 1970s, the skies grew clearer and sea surface temperatures increased.

In the end, Booth says, the changing output of industrial aerosols explains two-thirds of the long-term swings observed in sea surface temperatures in the North Atlantic. "It's only in the current generation of models that we can see that relationship physically," says Booth.

The Hadley Centre's results seem to overturn the prevailing wisdom in climate circles, which holds that the ups and downs in sea surface temperatures result from a natural ocean cycle dubbed the Atlantic multidecadal oscillation (AMO). Earlier research suggested that the cooler Atlantic temperatures associated with the AMO could have contributed to droughts over the Sahel in Africa during the latter half of the twentieth century; the same cooling effect may have led to a reduction in the force of tropical storms and hurricanes steaming towards America[2]. But on the basis of the new picture, human pollution could be causing these climate disruptions instead.

The question now is whether the results will hold up. Researchers at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, say that they see hints of similar effects in their new simulations. But not everybody is convinced that aerosol pollution could have such a profound effect on ocean temperatures — and consequently on climate. NCAR climate scientist Kevin Trenberth says that the results depend on uncertain estimates of aerosol pollution and cloud distributions around the Atlantic. At the same time, satellite observations do not find the indirect aerosol effect to be as strong as the models seem to suggest, he says. "It would be surprising to me if the ocean is not playing a substantial role" through natural cycles.



ARCTIC WARMERS

Researchers are also struggling to tease apart the roles of natural cycles and human-caused changes in the melting Arctic. The sea ice there has taken a beating during the past few decades and coverage reached a near-record low of 4.33 million square kilometres last September. Because the speed of the ice loss has outstripped all but the most dire model predictions, researchers have wondered what might be missing from their simulations.

Early results from the new models suggest that the addition of the more complex clouds and aerosols to simulations could help to provide an explanation. NCAR's new atmospheric model produced more warming and sea-ice loss than the previous iteration[3], and the culprit seems to be clouds — a result that caught researchers by surprise. "I'm a cloud girl, but I didn't go into this thinking that clouds were going to play the lead role," says Jennifer Kay, an atmospheric scientist at NCAR.

To figure out what was happening, the team

built new diagnostic tools into the model that effectively tell scientists what they would see if they were observing the planet from a pair of US satellites, CloudSat and Calipso. The model's output is translated into a signal that can be compared directly with radar and laser instruments aboard the satellites, Kay explains. "You basically fly a little satellite around inside the model," she says, "and what it shows is that the clouds are remarkably improved in the new version." They tend to be thinner and more transparent — more like their physical counterparts in the Arctic skies — although why remains unclear.

The gauzy clouds allow more sunlight through in the summer, which melts more ice and exposes more sea and land surfaces; these effects are enhanced by deposition of dark aerosol particles on the snow. It all adds up to a shift towards darker surfaces that absorb more sunlight and amplify warming. Although the model still tends to underestimate sea-ice loss on average, Kay says, some simulations lined up with satellite observations reasonably well.

Researchers at the GFDL are also seeing greater sea-ice declines with their new climate model. Michael Winton, a modeller at the GFDL, says this is likely to be a theme in the IPCC's fifth assessment, but he warns against premature celebration. The addition of enhanced clouds and aerosols to the simulations is driving the extra warming, but the exact details remain unclear[4].

In the end, the climate community must confront a basic question about models. "If you made a model and it matched the observations perfectly, would you claim success?" Winton asks. Although the new GFDL model has an enhanced representation of the atmosphere and does a better job of matching satellite observations, Winton warns that modellers could get the right answer for the wrong reasons. There is some evidence, for example, that natural variability in ocean circulation has caused some of the sea-ice loss during the past two decades. "The Arctic has to be understood in the context of the overall climate," he says.

TAMING THE MONSOON

In satellite images, southeast Asia is often covered by a giant blemish — a brown cloud fed by black carbon emissions from millions of primitive cooking stoves and open fires throughout rural India and neighbouring countries. In the atmosphere, those dark particles absorb sunlight and heat the surrounding air while cooling the land below, effectively stabilizing the atmosphere and slowing the regional circulation that draws moisture inland from the northern Indian Ocean. Researchers proposed seven years ago that this mechanism could explain why the south Asian summer monsoon has grown weaker over the past half-century[5].

However, simulations with one of the new models at the GFDL suggest that the situation might be more complicated, with aerosols and clouds disturbing a much larger hemispheric energy exchange[6].

The overall system is driven by the summer Sun, which delivers more heat north of the equator than south. In what amounts to a massive heat engine that redistributes energy between the hemispheres, hot air rises in the north and carries heat at altitude to the south, where the air descends and picks up moisture from the Indian Ocean on its return north. It is this last step that brings the summer monsoons, which provide up to 80% of the precipitation to most of India. But the GFDL results, reported in *Science* last October, showed that aerosols are creating a major disruption[6].

"Aerosol emissions are like putting up a sunscreen over the Northern Hemisphere, and that reduces the solar imbalance that drives the system," says Yi Ming, a GFDL climate modeller and an author of the study. "We're trying to argue this from a larger spatial scale."

Their model also shifts the blame away from the black-carbon emissions of cooking stoves and agricultural fires, and towards sulphur pollution from coal-fired power plants throughout the region. The sulphate particles that develop from such pollution serve as the seeds for water droplets and brighten clouds, cooling the land below. In addition to capturing the 4–5% overall decline in summer rainfall over India since 1950, the model reproduces regional variations in precipitation — more drying over north-central India versus a slight increase in rainfall over southern India and northwestern India and Pakistan. Ming says the indirect aerosol effect included in the new study shows "a different part of the puzzle".

Surabi Menon, a climate modeller and an affiliate scientist at the Lawrence Berkeley National Laboratory in California, cautions that the simulations rely on relatively incomplete estimates of emissions. Menon has been exploring aerosols and the monsoon with the latest model from the NASA Goddard Institute for Space Studies in New York, and says that modellers can at least check their data against measurements of pollution, which were not available even a few years ago. "We are getting there," she says. "Slowly."

THE GLOBAL PUZZLE

As climate researchers test drive the new generation of models, they are particularly keen to measure the models' overall sensitivity: how strongly they warm up in response to increasing concentrations of greenhouse gases. The addition of indirect aerosol effects makes the new model at NCAR more sensitive to greenhouse gases, says NCAR researcher Andrew Gettelman. Simulations show that the additional cooling from aerosol pollution, as well as the direct effect of haze, masked some of the warming from greenhouse gases during the twentieth century; but the model shows enhanced warming in the twenty-first century as curbs on pollution expose the full power of greenhouse gases. In simplified runs that double greenhouse-gas concentrations — which could happen by the end of this century — the new atmospheric model projects a 4 °C rise in global temperatures, whereas the previous model showed a 3.1 °C increase.

The Hadley Centre model is moving in the same direction, but this is not the rule. A model at the Pierre Simon Laplace Institute near Paris produces less warming in response to greenhouse gases than did the previous generation, says Sandrine Bony, a climate modeller there. The improved treatment of clouds may help explain that change, but the researchers have yet to fully analyse the new results.

These are just the first wave of a deluge in modelling data. Scientists in the IPCC's physical science working group have until 31 July 2012 to submit papers for the IPCC process, so the literature will explode with results from climate simulations over the coming year.

Then the real hard work will begin — working out what to believe. Scientists must tease apart the subtle causes and effects in their models and, where possible, test their results against other models and observations. "What we need now is to really understand what the models are doing, and why they differ," Bony says. "It's really by comparing the results from a spectrum of models that we can assess which results are robust." ■

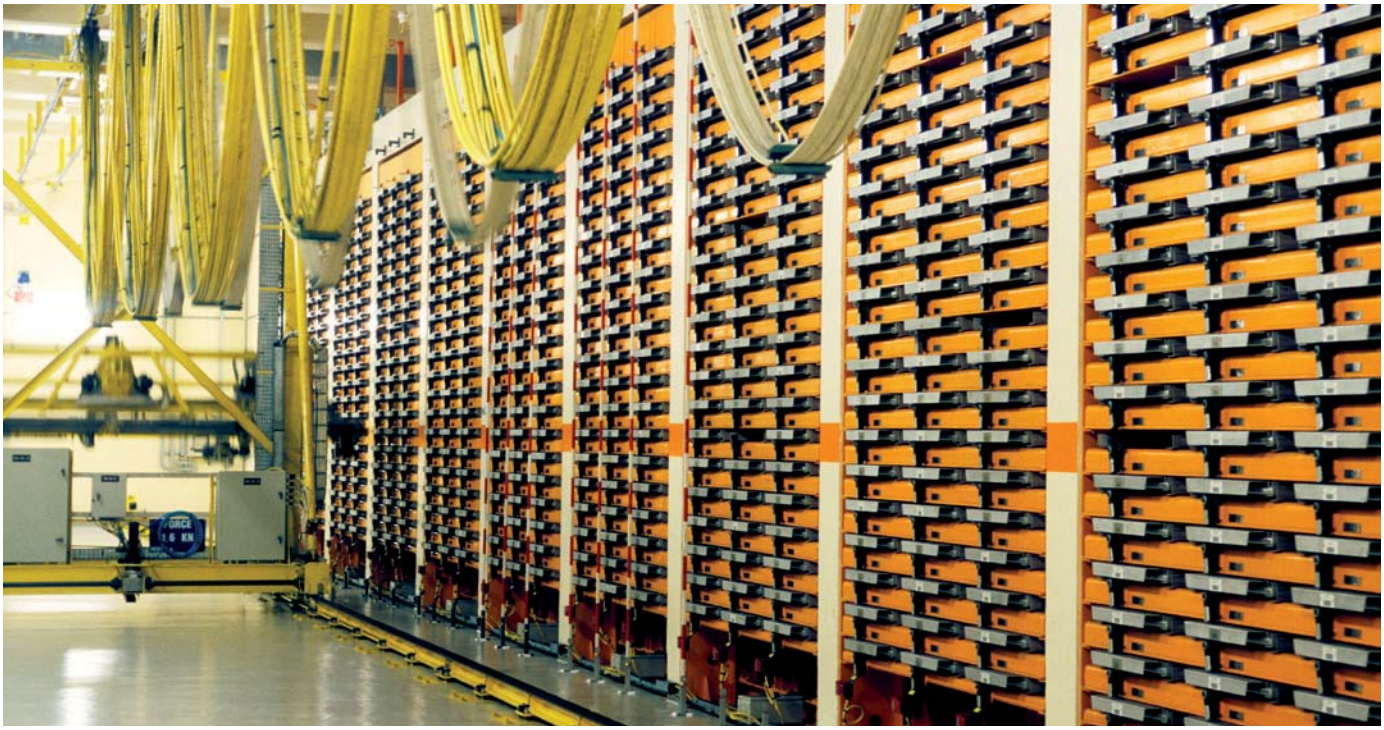
"What we need now is to really understand what the models are doing, and why they differ."

Jeff Tollefson *covers energy and environment for* Nature *in New York.*

1. Booth, B. B. B., Dunstone, N. J., Halloran, P. R., Andrews, T. & Bellouin, N. *Nature* http://dx.doi.org/10.1038/nature10946 (2012).
2. Mann, M. E. & Emanuel, K. A. *Eos Trans. AGU* **87,** 233–244 (2006).
3. Kay, J. E. *et al. J. Clim.* http://dx.doi.org/10.1175/JCLI-D-11-00622.1 (2012).
4. Winton, M. *J. Clim.* http://dx.doi.org/10.1175/2011JCLI4146.1 (2011).
5. Ramanathan, V. *et al. Proc. Natl Acad. Sci. USA* **102,** 5326–5333 (2005).
6. Bollasina, M. A., Ming, Y. & Ramaswamy, V. *Science* **334,** 502–505 (2011).

COMMENT

P. LANDMANN/SPL



The production of plutonium nuclear fuel in France (storage facility shown) adds millions of dollars each year to the cost of electricity generation.

Time to bury plutonium

Recycling plutonium is dangerous and costly. Britain should take the lead on direct disposal, say **Frank von Hippel**, **Rodney Ewing**, **Richard Garwin** and **Allison Macfarlane**.

The world today has a stock of about 500 tonnes of separated plutonium — enough to make 100,000 nuclear weapons[1]. As observed in a report[2] by the US National Academy of Sciences in 1994, this material is "a clear and present danger to national and international security". Yet, almost two decades later, programmes to dispose of it are in disarray.

This plutonium is a legacy of the cold war, and of a 1960s enthusiasm for a nuclear-powered future using revolutionary plutonium 'breeder' reactors. Some countries separated plutonium from the spent fuel of uranium-fuelled nuclear plants, expecting to use it to power this new generation of reactors. But the revolution never materialized. Only Russia (which has the world's largest plutonium stockpile, when counting both civilian and weapons stocks) and India still have active programmes to commercialize breeder reactors in the near term.

The United Kingdom, which owns the largest civilian stocks of separated plutonium (about 90 tonnes), announced plans last December to manufacture it into fuel for proposed water-cooled nuclear power reactors. This proposal matches that of the United States — to use already-separated plutonium as an alternative fuel for existing nuclear power reactors. France and Japan, the other nations with significant stockpiles, combine this approach with the dangerous and costly policy of continued separation of plutonium from spent fuel, which prolongs the associated international security risks.

On the basis of past experience in Britain, the United States and Japan, the UK strategy is likely to run into technical and political difficulties, as well as escalating costs. Before major investments are made, Britain should seriously evaluate the less costly and less risky method of direct plutonium disposal, and take the opportunity to lead the world towards a better solution for reducing stockpiles.

THE ALTERNATIVES

The 1994 report[2] by the National Academy of Sciences focused on two alternatives for plutonium disposal. The first involves mixing plutonium with depleted uranium ▶

to make 'mixed oxide' (MOX) fuel that can be used by current-generation nuclear power reactors. Once used, the MOX fuel needs to be disposed of with other spent reactor fuel. The second option is direct disposal: immobilizing the plutonium in ceramic and burying it in a geological repository with spent fuel or radioactive waste. Both options require about the same repository space[3].

In 1994, France was already pursuing the MOX fuel option as part of a larger, controversial strategy of separating and recycling plutonium from its spent uranium-based nuclear fuel. It initially separated plutonium for weapons, and then for demonstration breeder reactors. After becoming the global expert in this technology, France's government-owned nuclear services company, now called Areva, built a reprocessing plant to separate plutonium from the spent fuel of other countries. However, Areva's main foreign customers have not renewed their contracts, and the national electricity utility is becoming increasingly restive about having to support a domestic MOX programme that is making France's own power more expensive. According to a 2000 estimate, recycling plutonium from spent fuel adds about US$750 million each year to the cost of electric power generation in France, in comparison with the cost of using fresh uranium fuel and disposing the waste in a geological repository[4].

Japan has pursued a similar strategy of reprocessing spent fuel and using it in MOX, largely to put off a politically difficult decision about where to site a nuclear waste repository. It built its own costly reprocessing plant for domestic fuel — designed mostly by Areva — but escalating costs and delays prevented completion by more than a decade. The plant separated about 4 tonnes of plutonium in 2006–08, but was forced to shut down because of a malfunction. An attempted restart in January this year resulted in the same malfunction. Construction is scheduled to start this spring on a MOX fuel fabrication plant but, following the Fukushima accident in March 2011, Japan's entire nuclear programme is being reviewed.

"By 2018, Britain's stock of separated plutonium will have increased to more than 100 tonnes."

Britain's Nuclear Decommissioning Authority is now completing contracts to separate plutonium from UK spent reactor fuel. By 2018, when the two UK reprocessing plants are expected to have fulfilled their contracts, they will have increased the country's stock of separated plutonium to more than 100 tonnes. In December 2011, the UK Department of Energy and Climate Change tentatively concluded that the best option for disposing of this plutonium would be to buy a new MOX fuel fabrication plant.

Plutonium fuel pellets must be precisely made.

O. ANDERSEN/AFP/GETTY

LEARNING FROM HISTORY

Previous attempts to produce MOX in Britain have seen poor results. A fabrication plant at the Sellafield reprocessing site in Cumbria opened in 2001, initially to deal with plutonium separated for Japan. But because of design flaws and difficulties in achieving the exact manufacturing standards of MOX fuel, the plant operated at only 1% of its design capacity in its first ten years. In August 2011, after expenditures of £1.4 billion (US$2.3 billion), it was shut down.

In evaluating methods for plutonium disposal, Britain should also consider the experience of the United States. It decided to pursue both MOX and immobilization routes, estimating in 1999 that it would cost about $4 billion to dispose of 34 tonnes of its 85-tonne stockpile of weapons-grade plutonium. But Russia, which had also committed to disposing 34 tonnes of its own weapons plutonium, objected to immobilization because the plutonium could be made into weapons if it were recovered. This, along with the cost of paying for two different programmes, led the United States to abandon the immobilization track. Instead, it commissioned an Areva-designed MOX plant. The cost of disposing of its 34 tonnes of plutonium has since soared to more than $13 billion[5], with the value of fuel produced likely to offset costs by only $1 billion to $2 billion.

Britain should therefore give plutonium immobilization another look. Although the technique has not been demonstrated at full scale, there is substantial literature on how to do it[6,7]. Immobilization should be easier and cheaper than MOX production. Converting 100 tonnes of plutonium into MOX fuel requires fabricating 100 million pellets of fuel, machined to exact dimensions to fit into long zirconium tubes. For disposal, however, the plutonium could be immobilized in fewer, less-precisely-sized 'pucks'.

This immobilized plutonium could be packaged with spent fuel or solidified reprocessing waste, which emits γ-radiation that would ward off any thieves or terrorists for a century before its disposal in a 500-metre-deep geological repository. Another option would be irreversible disposal in a few 5,000-metre-deep boreholes. The National Academy of Sciences discussed this method in 1994, and there has been more design work since[8,9]. Britain's decommissioning authority found in a 2009 study that most immobilization options would be less costly than MOX but are technologically less mature (see go.nature.com/rbxmsl). The failure of the UK MOX plant and other problems, however, suggest that immobilization is lower risk.

The United Kingdom is ideally placed to spearhead this effort. It has the world's largest stockpile of separated civilian plutonium and has seen the failure of a MOX plant. It should take the lead in developing plutonium immobilization through laboratory-scale tests, a pilot project and then a full-scale plant. It is time to follow a different path, in which plutonium is treated unambiguously as the dangerous weapons material that it is. ■

Frank von Hippel *is professor of public and international affairs at Princeton University, New Jersey, USA, and co-chair of the International Panel on Fissile Materials.* **Rodney Ewing** *is professor of earth and environmental sciences at the University of Michigan, Ann Arbor, USA.* **Richard Garwin** *is a physicist and IBM Fellow emeritus at the Thomas J. Watson Research Center, New York, USA.* **Allison Macfarlane** *is associate professor of environmental science and policy at George Mason University, Fairfax, Virginia, USA.*
e-mail: fvhippel@princeton.edu

1. International Panel on Fissile Materials. *Global Fissile Material Report 2011* (IPFM, 2011); available at http://go.nature.com/fi58zk
2. Committee on International Security and Arms Control. *Management and Disposition of Excess Weapons Plutonium* (National Academies Press, 1994).
3. Wigeland, R. A. *et al.* Paper 496 in *Proc. GLOBAL '05*, Tsukuba, Japan, October 2005.
4. Charpin, J. M., Dessus, B. & Pellat, R. *Economic Forecast Study of the Nuclear Power Option* (Commissariat Général du Plan, France, 2000).
5. US Department of Energy *FY 2013 Congressional Budget Request* Vol. 1, 460–461 (2012); available at http://go.nature.com/4posor
6. Yudintsev, S. V. *et al.* in *Structural Chemistry of Inorganic Actinide Compounds* (eds Krivovichev, S. V. *et al.*) 457–490 (Elsevier, 2007).
7. Ewing, R. C. & Weber, W. J. in *The Chemistry of the Actinides and Transactinide Elements* Vol. 6 (eds Morss, L. R. *et al.*) 3813–3887 (Springer, 2011).
8. Halsey, W. G., Jardine, L. J. & Walter, C. E. *Disposition of Plutonium in Deep Boreholes* (Lawrence Livermore National Laboratory, 1995).
9. Gibb, F. G. F., Taylor, K. J. & Burakov, B. E. *J. Nucl. Mat.* **374,** 364–369 (2008).

E. ZURGA/BLOOMBERG NEWS/GETTY

US sales of dietary supplements exceed US$28 billion a year, but ingredients are unregulated.

Strengthen and stabilize the FDA

The US Food and Drug Administration needs to be more independent, says **Daniel Carpenter**.

There is perhaps no more important public-health agency in the world than the US Food and Drug Administration (FDA). Its policies have reshaped science and regulation worldwide, giving billions of people greater confidence in the treatments and foods on which they rely[1]. Yet the agency's capacity and autonomy — and hence the services it renders — are in jeopardy.

The FDA is plagued by threats to its power and stability. A vivid example of this came last December, when the agency was shockingly overruled by Kathleen Sebelius, secretary of the Department of Health and Human Services (DHHS). She decided, with public backing from President Barack Obama, that the contraceptive drug Plan B would not be available to teenagers under the age of 17 without a prescription.

With this move, Sebelius quashed an eight-year decision process, turned back more than 70 years of precedent in which the agency's decisions are final, and invited future drug-approval contestants to take their case directly to the White House.

At the same time, the agency has little jurisdiction over a growing segment of the health-care system: dietary supplements. This regulatory gap has had deadly consequences. Today, dozens of athletic supplements sold throughout the United States contain DMAA (1,3-dimethylamylamine), a stimulant similar to amphetamine that was withdrawn from the US pharmaceutical market in the 1970s because of health concerns.

In 2010, US sales of supplements containing DMAA exceeded US$100 million. DMAA has been linked to increased blood pressure and heart rate, panic attacks, seizures and stress-induced cardiomyopathy. After two deaths last year, the US military in December stopped the sale of supplements containing DMAA on its military bases. Last summer, Health Canada banned DMAA from all supplements.

Amid such developments, many things have been going well for the FDA. The agency approved a near-record number of medicines last year[2], and it brings new cancer therapies to market quicker than its counterpart, the European Medicines Agency[3]. And the FDA has asserted its independence at times — under immense pressure to continue permitting the drug bevacizumab (Avastin) to be marketed for metastatic breast cancer, the agency instead followed the scientific evidence and revoked its approval in November 2011. (Avastin remains available for off-label prescription and for other cancers.)

The agency has also demonstrated strength and flexibility in its regulation of diet pills. It removed sibutramine (Meridia) from the markets and did not approve rimonabant (which was approved, then withdrawn, in Europe), but it has been willing to consider new evidence for the diet pill Qnexa (a mixture of phentermine and topiramate).

Still, the FDA's recent misfortunes leave room for concern. They come at a difficult time for US science, society and politics, during which the country's health sector has grown weaker. Until Congress acts to boost the FDA's strength and independence, the safety and confidence of the world's citizens — as well as medical and technical innovation — remain at risk. I propose a series of realistic reforms; they are not a panacea for the FDA or for US public health, but they could help to preserve the FDA's place as the pre-eminent regulatory agency in the world.

A STRONGER BODY

The priority in any reform is to strengthen the agency. As a first step, we should make the FDA a truly independent body. We should separate it from the DHHS and give the FDA commissioner a six-year term like that of the chair of the US Federal Reserve, to be deposed only 'for cause'. Agency responsibilities should be transferred from the DHHS secretary to the FDA commissioner, which would prevent future repeats of the Plan B events by placing all drug-approval decisions in the hands of the FDA, not the White House.

In addition, we should reform how the agency is funded. At present, the FDA is partly supported by application fees that drug companies pay each time they submit a new drug for approval. The rates and terms of these fees — or, more appropriately, taxes — are renegotiated every five years, creating an opportunity for agency critics to hold up funding until their demands are met, destabilizing drug development and consumer protection.

Negotiations with companies are conducted in secret, with citizens and safety advocates effectively excluded, and research has shown that drugs approved just before the drug-review deadline are more likely to encounter safety problems. The list of drugs that were approved under deadline pressure and then pulled from the market

J. REED/REUTERS/CORBIS

The FDA, headquartered in Silver Spring, Maryland, approved a near-record number of drugs last year.

because of dangerous side effects includes Vioxx, Bextra, Rezulin, Baycol, Trovan and Avandia (withdrawn in Europe)[4,5].

Instead of taxing new drug submissions to fund the FDA, the US government should tax the thing that benefits from the FDA's backing — pharmaceutical sales. The FDA induces great confidence in the nation's drug supply; the drug companies and citizens who benefit from that confidence should pay a small amount of revenue for it. To avoid giving the FDA incentive to approve and facilitate only blockbuster medicines, Congress could put a cap on the revenue raised from such a tax.

GREATER FLEXIBILITY

The agency's power over herbal and nutritional supplements should be strengthened. The DMAA example is not unique; every year, people in the United States spend more than $28 billion on supplements with the (mistaken) presumption that they are safe and effective. However, according to the 1994 Dietary Supplement Health and Education Act, supplements are not required to be proven effective, and any supplements on sale since before 1994 are assumed to be safe.

As Pieter Cohen, a dietary-supplements expert at Harvard Medical School in Boston, Massachusetts, has suggested[6], the United States should require testing of all dietary supplements. Exceptions should be made only for those substances that are 'generally recognized as safe'.

To strengthen drug development, we should reduce its cost for certain types of disease. I propose that we create a quicker, conditional approval system for drugs aimed at treating illnesses that are especially lethal or that constitute genuine public-health crises. This would extend and strengthen what the FDA and Congress have done for AIDS and cancer. For the diseases that public-health officials consider to be the deadliest and most undertreated (such as gastric and lung cancers, neurodegenerative diseases and some infectious diseases), phases II and III of the drug-development process could be more systematically merged.

The FDA could approve new drugs for these illnesses for an initial five-year window, with follow-up studies required before the drug could be re-authorized for another five years. This would create stronger incentives for companies to conduct post-market trials, and would push more of the high costs of phase III trials to the post-market phase, where they could be more easily covered by sales revenue.

"To fund the FDA, tax the thing that benefits from the FDA's backing — pharmaceutical sales."

Another way to facilitate drug development is to have the FDA open its data vaults. Pharmaceutical companies spend too much time chasing drug ideas that someone else has already tried, found to fail and abandoned. The FDA receives information on every clinical trial conducted, but it is prohibited from sharing data on failed drugs, and companies don't have unilateral incentives to share data.

With a modest but crucial change in federal statute, we could have a system in which data on failed drug-development projects are shared with companies and citizens, while the most sensitive information, such as how a company uncovered a biological or pharmacological mechanism of action, is kept proprietary. Knowledgeable sources in the pharmaceutical industry tell me that this step alone could reduce costs by 10–20%. Sharing data on abandoned drugs would also help to improve drug safety, by showing which drugs (and which substances and mechanisms of action) encountered safety and toxicity problems in the experimental stage.

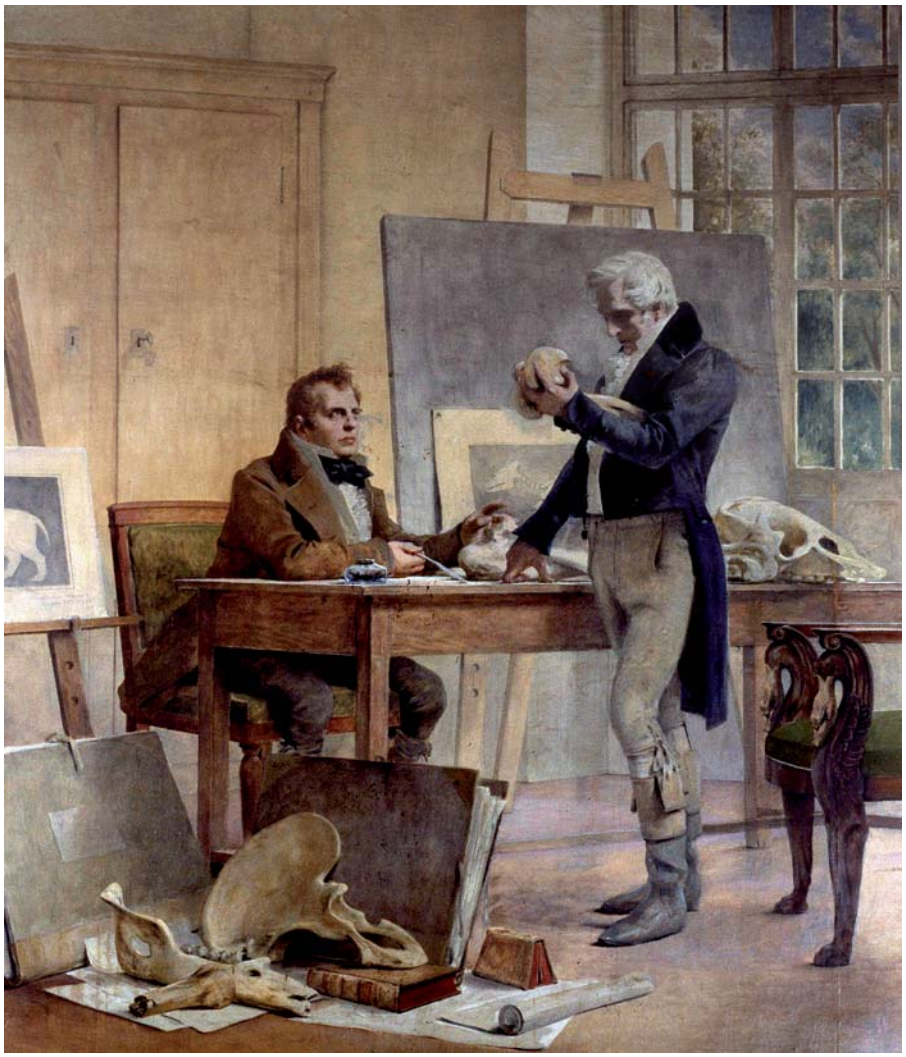
TWENTY-FIRST-CENTURY AGENCY

To improve the public's trust in the FDA, Congress should set deadlines for the commissioner's decisions, which have been slow in recent years. When ruling on Avastin and the diabetes drug Avandia, the commissioner took at least three months to make a final decision after the vote from the advisory committee. This kind of delay looks bad, not least because the scientific opinion has been clarified, making any lag seem to be caused by the worst form of politics. Once the advisory committee has voted on an issue, the commissioner should have no more than a month to make a decision.

None of these reforms can be accomplished by the FDA alone; almost all of them would require changes to statute that can come only from the US Congress. And this gets us to the heart of the problem. Lately, the toxic polarization between Democratic and Republican lawmakers — and the micromanaging tendencies of Barack Obama and former President George W. Bush[7] — are damaging US public-health infrastructure and its scientific prospects, weakening one of the republic's most vital institutions. A robust FDA for the twenty-first century demands selective strengthening of the agency and flexibility on key dimensions. ■

Daniel Carpenter *is Allie S. Freed Professor of Government at Harvard University, Cambridge, Massachusetts 02138, USA, and a visiting researcher at the Institute for Political Studies at the University of Strasbourg, Strasbourg F-67081, France. e-mail: dcarpenter@gov.harvard.edu.*

1. Carpenter, D. *Reputation and Power: Organizational Image and Pharmaceutical Regulation at the FDA* (Princeton Univ. Press, 2010).
2. US Food and Drug Administration. http://go.nature.com/gtnheq (2011).
3. Roberts, S. A., Allen, J. D. & Sigal, E. V. *Health Affairs* **30,** 1375–1381 (2011).
4. Miller, J. D. *J. Am. Med. Assoc.* **302,** 189–191 (2009).
5. Carpenter, D., Chattopadhyay, J., Moffitt, S. & Nall, C. *Am. J. Polit. Sci.* **56,** 89–114 (2012).
6. Cohen, P. A. *N. Engl. J. Med.* **361,** 1523–1525 (2009).
7. Harris, G. White House and the FDA Often at Odds. *New York Times* (3 April 2012).



T. CHARTRAN/WHITE IMAGES/SCALA, FLORENCE

Georges Cuvier (standing) was one of dozens of naturalists who laid the groundwork for Charles Darwin.

NATURAL SELECTION

The evolutionary struggle

Andrew Berry enjoys a biographical feast that turns the spotlight onto Darwin's forerunners.

It is remarkable that the theory of evolution has come to be associated exclusively with Charles Darwin. Even Alfred Russel Wallace, co-author of the paper that first unveiled evolution by natural selection, has mostly disappeared from view. In *Darwin's Ghosts*, novelist and science historian Rebecca Stott explores the intellectual origins of the theory of natural selection through scientific biographies of Darwin's antecedents and contemporaries, from Aristotle to Wallace.

The usual suspects are here, including French naturalists Jean-Baptiste Lamarck, Georges Cuvier and Georges-Louis Leclerc, Count of Buffon. But so are people whose contributions to the history of evolutionary theory are generally known only in history of science departments, such as Swiss biologist Abraham Trembley and French natural historian Benoît de Maillet. Stott's research is broad and unerring; her book is wonderful.

➲ NATURE.COM
For more on Alfred Russel Wallace, see:
go.nature.com/s9z8op

Darwin's Ghosts: In Search of the First Evolutionists
REBECCA STOTT
Bloomsbury/Spiegel & Grau: 2012. 400/416 pp. £25/$27

On the Origin of Species (John Murray, 1859) was rushed out. In June 1858, Darwin got a letter from Wallace, then in Indonesia, suggesting the idea — evolution by natural selection — that Darwin had been quietly gestating for 20 years. Only intervention by colleagues saved Darwin's claim to precedence. The outcomes were a paper co-published by Darwin and Wallace in the *Journal of the Linnean Society* in July 1858, and *Origin* in November the next year.

After the publication, Darwin's materialistic vision of biological change was, as he had feared, condemned as heretical. But blasphemy was not the only charge laid at his door: some of Darwin's correspondents complained that he had plagiarized their work.

Darwin saw *Origin* as a quick and dirty synopsis of his ideas, not the planned 'big species book', as he referred to it. One casualty was a review of the literature. As Stott recounts, Darwin dealt with this oversight (and the critical letters) in 1861, by adding a review, *An Historical Sketch of the Recent Progress of Opinion on the Origin of Species*, to the third edition. Stott's book presents encounters with the inhabitants of this addendum, plus a few who did not make Darwin's cut.

The *Sketch* was an honest attempt to give credit where it was due. But it is clear that Darwin was keen, by omission, to emphasize his own claim to the theory. Wallace is mentioned just four times in the 490 pages of the first edition of *Origin*. And in his autobiography, Darwin downplayed the influence of his grandfather, Erasmus Darwin, whose evolutionary speculations were both historically significant and part of his family's lore.

In looking beyond Darwin, Stott deals with eye-wateringly complicated material. A three-way chapter on Lamarck, Cuvier and fellow French naturalist Étienne Geoffroy, for instance, describes — with a novelist's eye for dramatic detail — how, in the early nineteenth century, they jockeyed for pre-eminence at the newly formed French National Museum of Natural History in Paris.

More than the story of three careers, this is also about the waxing and waning of friendships, a clash of deeply opposed world views and some of the most exciting and innovative science ever done. And the story is complicated by difficulties in interpreting the documentary record, which is mostly a monument to courtesy. Cuvier long suppressed his unfavourable view of Lamarck, waiting instead to bury both Lamarck's ideas and their author ▶

with a single brutal obituary, published in the *Memoirs of the Royal Academy of Sciences of the Institute of France* in 1835.

Stott highlights the charged moment when Cuvier first examined mummified ibises collected by Geoffroy on the Napoleonic expedition to Egypt. Here was the ultimate showdown between Lamarck's evolutionary ideas, which predicted that ibises should have experienced species change in the 3,000 years since the specimens were alive, and Cuvier's insistence that this was biologically impossible. Were the ancient ibis mummies significantly different from modern birds? No — Cuvier seemed to have been proved right.

Many of the heroes of *Darwin's Ghosts* ran risks to pursue their evolutionary ideas — in 1749, for example, French philosopher Denis Diderot was imprisoned for subversive writings that touched on species variation. Many thinkers tried to sidestep the charge of heresy: de Maillet, for example, distanced himself by presenting his theories in the form of a supposed conversation with an Indian mystic, 'Telliamed' (de Maillet spelled backwards). Erasmus Darwin, anxious about the impact of controversy on his reputation as a doctor, chose to veil many of his evolutionary speculations behind a cloak of classics-tinged poetry. Scottish geologist Robert Chambers never publicly admitted that he was the author of the anonymous Victorian best-seller *Vestiges of the Natural History of Creation* (John Churchill, 1848).

The lesson of Stott's book is that Darwin and Wallace were not just standing on the shoulders of giants scientifically. They were also at liberty to speculate and publish freely on the topic only because of the risks that these earlier writers had taken.

Stott introduces us to a sparkling cast of characters, but the biographical approach has its limitations. The book fails to illuminate the most remarkable aspect of the story of the discovery of evolution: that this long-sought-after idea was discovered independently, around the same time, by two men who both regarded themselves as pedestrian thinkers.

The Darwin–Wallace story validates the modern insistence that discovery is not about 'great men', but about a confluence of societal and technological factors that collectively make a previously inaccessible idea accessible. Nevertheless, Stott's constellation of biographies is an exhilarating romp through 2,000 years of fascinating scientific history. ■

Andrew Berry *is a lecturer in evolutionary biology and teaches history of science at Harvard University in Cambridge, Massachusetts.*
e-mail: berry@oeb.harvard.edu

SNIBBE INTERACTIVE

Social Light **at the Science Museum in London lets audience members act as mirrors.**

Q&A Scott Snibbe
Nature's digitizer

Media designer Scott Snibbe creates software apps and interactive science-museum installations, and was executive producer of the 2011 Biophilia *project by singer–songwriter Björk. As he prepares to lecture at the Sónar International Festival of Advanced Music and New Media Art in São Paulo, Brazil — where his visuals will accompany Björk's performance of* Biophilia *— he talks about provoking wonder.*

How did you become a digital designer?
My father is an inventor who designed a geometric kite and is working on a perpetual-motion machine. My mother is an artist. From childhood I wanted to be a combination of the two. My parents let me use a machine workshop from the age of four to make anything, however useless. My dad and I built a Tesla coil, and I got a few 20,000-volt shocks, but my parents weren't afraid because we were Christian Scientists, and didn't believe that God would allow us to get hurt as long as we had a positive attitude and safety goggles. At university, I considered studying genetics or neuroscience, but I couldn't handle dissections or vivisections. Instead, I became a researcher at the computer–human interface, working on problems such as artificial touch and computer vision at places including Brown University in Providence, Rhode Island, and Adobe Systems in Seattle, Washington. Then I created my own companies, combining interactive art with business.

What draws you to interactive apps?
Other fields are limited by money, equipment and the laws of nature. But with computers, the only limits are technical ability, ingenuity and imagination. Nature has awed me since I was a child, but the educational system rarely conveys this wonder, transforming our Universe into boring multiple-choice questions. My programs recreate the wonder and magic to give people the kind of experiences that they have in wild places such as river banks. My apps borrow from nature, but the laws are slightly altered, as if in a parallel universe.

NATURE.COM
For an interview with Bjork on *Biophilia*, visit:
go.nature.com/18n54b

NICK HIGGINS

Sónar International Festival of Advanced Music and New Media Art
Anhembi Parque, São Paolo, Brazil. 11–12 May 2012.

Can you describe your science-based apps?
With my *Gravilux*, you touch the screen and stars are attracted to your fingertips. I started with Newton's gravity equations but didn't get controllable patterns, so I removed mutual attraction. *Bubble Harp* draws Voronoi diagrams, based on a geometric algorithm first described by seventeenth-century philosopher René Descartes, and used to model the structure of cells, the pattern of human settlements and the gravitational influence of stars. With *Antograph*, you 'paint' a pheromone that attracts ants, but they swarm off the trail, just as real ants would. I've had reports of it being used to teach what pheromones are, and one user of *Gravilux* said that it helped him to get an A grade in physics for the first time.

How did you come to work with Björk?
Björk chose to release *Biophilia* on the iPad. She asked my studio to produce the

project, and to design several of the interactive song apps. One explains how viruses work: you see them injecting RNA into a cell and hijacking its reproductive mechanism. You can flick the viruses away, but if you do, the music stalls; you have to let the cell be attacked to hear the whole song. Another app, *Hollow*, animates DNA replication using a drum machine. When you touch different enzymes, they catalyse the DNA strand and trigger gothic musical loops.

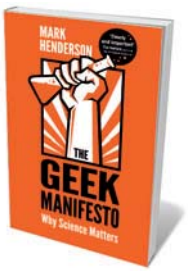
What makes a good science exhibit?
It must satisfy someone with a PhD — and a two-year-old. *Social Light*, an exhibit on electromagnetism that I designed for the Science Museum in London, allows your body to refract simulated light like a prism, reflect it like a mirror or absorb it. At the Exploratorium in San Francisco, California, *Three Drops* shows how forces of nature work at different scales. There is a screen where you can take a virtual shower as water flows around your shadow. Then the image zooms in to a single drop, which you can bounce around; the surface tension is so strong you can't get 'wet'. Then it zooms further in to show water molecules attracted to people's bodies as if they were impurities in the water. Here, we drew on the work of molecular biologist Tanya Raschke, who showed that water molecules form chains and loops.

How does the world of science differ from that of art?
There is an irreproducible uniqueness to an artist's work that makes the field less stressful than science. In science, if you don't make a certain discovery, someone else will, so even people in the same lab are competing with one another. In art, innovation and risk-taking are lauded, but in science there is an aversion to risk because people need to get grant money from conservative review boards. I know scientists who could speak a single sentence that would completely ruin their careers. And, like Barbara McClintock's pioneering work describing genetic crossover in corn, that sentence might even be true a decade later.

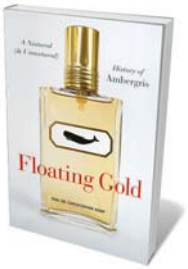
What keeps you excited about your work?
My imagination can take me up to Jupiter, or down to the size of atom — there is no need to actually create something unless it's for an audience. That is why I have mostly stopped showing in art galleries, because I wanted to reach the general public. I try to make an interactive app or exhibit as perfect as it can be, and then release it to see how people respond. I feel satisfied when someone says that our work was the most wonderful thing they encountered in their day. ■

INTERVIEW BY JASCHA HOFFMAN

Books in brief



The Geek Manifesto: Why Science Matters
Mark Henderson BANTAM *336 pp. £18.99 (2012)*
A geek revolution is upon us, asserts journalist Mark Henderson. Media stars such as physicist Brian Cox have lit the fuse by giving science cultural credibility. Now, says Henderson, with 7% of the UK electorate engaged or trained in science and more than 5 million scientists and engineers working in the United States, this sector of society is poised to gain real political clout. Ultimately, he argues, that could force change in everything from politics and government to health care and the environment as the intellectual honesty and innovative bent of the scientific mindset gains ground.



Floating Gold: A Natural (and Unnatural) History of Ambergris
Christopher Kemp UNIV. CHICAGO PRESS *232 pp. $22.50 (2012)*
Costly it may be, but the perfume fixative ambergris is weird stuff: a waxy mix of secretions and squid beaks from the intestines of the sperm whale. As molecular biologist Christopher Kemp relates, the beaks pass through the beast's four stomachs to form a dung-drenched mass. Often released when the whale dies, the floating lumps are seasoned by sea and sunlight, developing an odour likened to sandalwood, Brazil nuts and violets. Kemp's engrossing study takes us through history, tales of present-day hunters and cetacean science, poking its nose into the perfume industry on the way.

The Universe in Zero Words: The Story of Mathematics as Told Through Equations
Dana Mackenzie PRINCETON UNIV. PRESS *224 pp. $27.95 (2012)*
Mathematician and writer Dana Mackenzie brings to life 24 of the great equations that shape our world. We get Brahmagupta's subtle discovery of zero in 628 AD, the 350-year conundrum of Pierre de Fermat's last theorem, speculation over whether apples or moons inspired Isaac Newton's laws, the economic Black–Scholes formula that failed to prevent the Wall Street meltdown — and much more. Quietly learned and beautifully illustrated, Mackenzie's book is a celebration of the succinct and the singular in human expression.

Mr. Hornaday's War: How a Peculiar Victorian Zookeeper Waged a Lonely Crusade for Wildlife That Changed the World
Stefan Bechtel BEACON *272 pp. $26.95 (2012)*
One-time taxidermist William Temple Hornaday emerges from this lively biography as a nineteenth-century conservation hero — and a rampant racist. Stefan Bechtel tells how Hornaday saved the American bison and fought for legislation to save threatened species. Yet in 1906, as director of the Bronx Zoo in New York, he displayed Congolese pygmy Ota Benga in a cage, despite the protestations of local black clergymen. A fascinating portrait of a man both ahead of his time, and deluded by gross misreadings of Darwin.

Trusting What You're Told: How Children Learn from Others
Paul L. Harris HARVARD UNIV. PRESS *266 pp. $26.95 (2012)*
Children get information in two ways: from their own observation and exploration, and from other people. When the streams conflict, says educationalist Paul Harris, children often defer to the suggestions of others. But they are not uncritical: they "monitor the messenger", choosing whom to believe. Harris's challenge to the view of children as mini-scientists in a world-as-lab is well backed by research: a gripping trawl through the young human mind confronted with moral reasoning, the separation of fact from fiction and more.

Correspondence

Indian science needs alternative investors

India needs to look beyond its government and find other investors that can transform the face of science there (*Nature* **484,** 159–160; 2012). The private sector and philanthropic organizations must step in if the country is to realize its huge scientific potential.

India has opened up its financial markets, telecommunications and retail to foreign investment — education and science should follow suit. It should provide incentives for foreign universities, funding agencies and multinational corporations to invest in India's education and research, within a regulatory framework under strict oversight. The government could help to promote the economic and social returns on these investments to other interested parties.

India's nouveau riche need to follow the example of US and UK philanthropists and set up research foundations. Together with non-profit organizations, they can invest in high-risk, low-return areas of science not usually considered by the private sector.

As a young Indian scientist who aspires to build up a public-health genomics research programme in India, I believe that input from a variety of funding sources will help to drive competition and improve scientific output and accountability.

Siddhartha P. Kar *University of Texas, Houston, Texas, USA. siddhartha.p.kar@uth.tmc.edu*

Cyprus Institute: it deserves more credit

Your pessimistic report on the state of the Cyprus Institute in Nicosia (*Nature* **484,** 14; 2012) is based on a selection of distorted negative quotes. As chairman of the institute's trustees, I believe that this view is misleading and damaging to its hard-won scientific reputation in a region with little research history.

The morale of institute staff is low at present, but that is down to repeated delays in state funding rather than to its president's management style. You highlight harsh criticism from Nicholas Papadopoulos, chair of the parliament finance committee, but he has never visited the institute — despite repeated invitations. His comments are at odds with the government's whole-hearted support (see go.nature.com/fkqphb).

You are incorrect in saying that the institute signed an agreement with the Massachusetts Institute of Technology (MIT) to design a solar-energy plant, which is an initiative funded by the European Union; MIT is one of the partners in the institute's Energy, Environment and Water Research Center, which, among other projects, is in charge of the plant. You should also have mentioned that an audit found no improprieties in the institute's finances.

In our interviews for the story, Jos Lelieveld and I expressed full confidence in the institute and its management, but our views failed to come over in your report. I stressed, among other things, the board and scientific council's overwhelming acknowledgement of the institute's development and successes, and of the credit due to its staff and leadership, particularly its president.

Edouard Brézin *Laboratory of Theoretical Physics, Ecole Normale Superieure, Paris, France. brezin@lpt.ens.fr*

Cyprus Institute: improve oversight

As a scientist formerly employed by the Cyprus Institute, I believe that it is time for the Cyprus government to bring in new management that will have proper oversight (*Nature* **484,** 14; 2012).

The government auditor's report on the institute's finances was unfortunately published only in Greek. It highlights several irregularities, including a lack of transparency and formal organizational structure.

Allocation of resources has been inadequate and several staff members, including myself, have found the institute a difficult place to do research. In 2010, 13% of the staff resigned, and 20% of the remainder left in 2011 — including some senior staff members. This has been bad for research in Cyprus.

Considering the amount of money the government has poured into the institute and the number of researchers on the staff, the institute's scientific output seems low — and some of the publications listed on its website do not carry the institute's affiliation.

An independent committee should be formed to evaluate the institute's research according to international standards. This would replace its scientific advisory council, which consists mainly of members of its board of trustees and scientists from affiliated institutions.

Emmanouil Lioudakis *Cyprus Physicists Society, Nicosia, Cyprus. manolis.lioudakis@gmail.com*

A UK tissue bank for breast tumours

As chairman of the Breast Cancer Campaign Tissue Bank Management Board, I, along with my co-signatories, believe that researchers should have better access to breast-cancer tissue. Christina Curtis and her colleagues, for example, had to approach five tissue banks for the 2,000 samples they used to identify ten distinct types of breast tumour (C. Curtis *et al. Nature* http://doi.org/hvk; 2012).

Historically, inaccessibility of tissue samples and materials for breast-cancer research has been a major obstacle to translating science into new treatments, with researchers sometimes spending months tracking down suitable samples. A shortage of good-quality tissue with matching clinical data has been another hindrance.

The multicentre UK Breast Cancer Campaign Tissue Bank was opened earlier this year to help solve these problems (see go.nature.com/swbsrj). By offering annotated samples to all breast-cancer researchers in the United Kingdom and Ireland, it will help to speed the translation of findings into benefits for patients.

Alastair M. Thompson *on behalf of ten co-signatories*, Dundee Cancer Centre, University of, Dundee, UK. sgriffiths@breastcancercampaign.org*

**See go.nature.com/xfckko for a full list.*

Science sociology began before Kuhn

David Kaiser marks the 50th anniversary of the publication of Thomas Kuhn's best-selling *The Structure of Scientific Revolutions* (*Nature* **484,** 164–165; 2012). It is only fair to point out that many of the same ideas had already been formulated by the Polish microbiologist and philosopher of science Ludwik Fleck in his 1935 study *Genesis and Development of a Scientific Fact* (translated into English in 1979). Kuhn acknowledged Fleck's contribution in the foreword to the first edition of his book, but this was ignored in the intense debate that followed its publication.

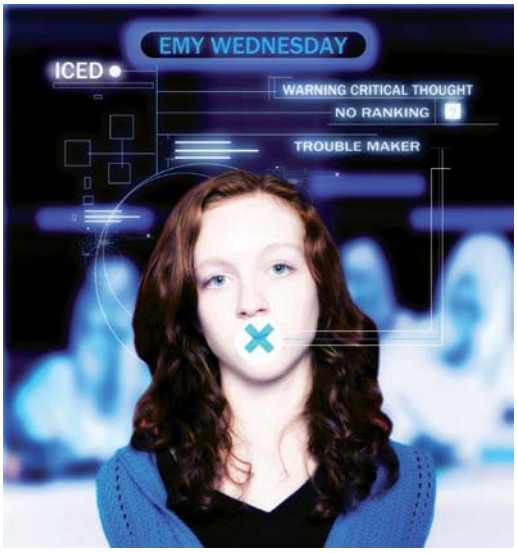
Fleck coined the term 'incommensurability' in 1927, which is still indispensable in discussions on Kuhn and the sociology of scientific knowledge. More on Fleck's theories and his influence on Kuhn's thinking can be found in the Stanford Encyclopedia of Philosophy (see go.nature.com/hpwnvd).

Ulrich Lehmann *Institute of Pathology, Hanover Medical School, Germany. lehmann.ulrich@mh-hannover.de*

THE APPROPRIATE RESPONSE

Class war.

BY JEFF SAMSON



JACEY

The last line rolled off the teacher's tongue. He turned from the wall displaying the poem and a portrait of its author, and looked at his students.

His eyepieces engaged immediately. Above a scattering of raised hands, each student's name burst into the air in bold, translucent blue characters.

As mandated, he ignored the customary hands, giving the less impetuous students ample processing time. He scanned the room, eyes moving from seat to seat, his gaze triggering the name hovering above each head to lift towards the ceiling, leaving a detailed profile in its wake.

"Mr Papillon," he said, and paused.

John Papillon — a socially challenged Level 3 with anger issues, who was never to be corrected or forced to maintain eye contact for longer than 6.5 seconds — shifted in his seat. Above his head, the details of his Individualized Cognition Enhancement Dossier — his ICED — were pushed left by a selection of questions that slid into position from empty space. The teacher chose the third.

"Mr Papillon," he read, "would you please be so kind as to recite the second word in the third line of the fourth stanza for us?"

The student squinted up at the poem, counting down and over with an outstretched finger, his tongue extended with effort.

"Home," he said.

The teacher turned and read the second word in the third line of the fourth stanza — hope. He turned back to John Papillon, as the profile and list of questions both slid left to be replaced by directions on how to respond to this student if — as was likely — he had misread his word. These too, the teacher read.

"Mr Papillon, you have correctly given us the third word of the fourth line of the second stanza. It is indeed 'home'. And it is entirely understandable that you selected the word you did, seeing as I threw quite a bit of information at you at once."

He paused as the warning light flashed, telling him to look away. He broke, then resumed eye contact.

"And on top of that, the word 'hope' is extremely close to the word 'home'. In fact, there is only one letter that makes them different. Well done, Mr Papillon."

He smiled for a calculated interval at the student, who stared back at him through glassy, half-lidded eyes. When the teacher turned away, the menus above John Papillon collapsed and his name fell back in line.

The teacher paused at Heidi Devendorf — a timid Level 4 with an acute fear of vowels, who was never to be asked to read aloud. A deep red light framed her profile, indicating that she'd already filled her allotted participation time for the week.

He stopped next at Seth Nozarski — a parentally inflated Level 8, who was to be kept at a minimum distance of seven feet from any student with freckles. But the violet glow rimming his profile told the teacher Seth's self-assuredness was running unhealthily low, and that he'd be best left alone.

He continued his scan and eventually called on Arthur Fink — an artistically inclined Level 15 with a flair for abstract thinking, who was entitled to submit illustrations in lieu of essays and exams.

"Mr Fink," he said. The student jolted upright in his seat. "What do you think the title, *Crossing the Bar*, means?"

The student craned his head upwards, closing his eyes. He then crouched over his desk and began furiously scribbling with his stylus on his EdPad. Two minutes later, the student slammed down his stylus and spun the EdPad around to face the teacher. What he displayed, left to right, was a crude sketch of a cross, followed by "+ ing + the +" in sloppy letters. Next to this was a rudimentary façade of a building bearing the name 'Murphy's Pub' above a doorway, in which stood what seemed to be a bearded old man vomiting onto the pavement.

Before the response slid into position, a third student, whose waving hand the teacher had steadfastly ignored, broke the silence.

"Bullshit!" she cried.

The teacher whipped his head around. He stared into the smoldering eyes of Emy Wednesday — a troublemaker whose cognitive deficiencies eluded numerical ranking but indicated an inclination towards critical thought, and an utter disregard for the self-esteem and unfairly distributed abilities of her peers. Her hands were pressed to her desk, balled up in tight fists.

"Total bullshit!"

The teacher stood, silent, gazing at the space above her head, waiting for a response to appear. But she still kept screaming.

"The title is a metaphor for death! The whole poem's about death!"

He waited.

"It's about an old man accepting his death and not wanting those he leaves behind to mourn him when he's gone!"

And waited.

"And you're going on about naming this goddamn word in this goddamn stanza!" she cried, jabbing her finger into the charged air between them. "And —"

They burst into the room, two disciplinarians clad in black, weaving through the desks, seizing Emy as she flailed and fought, dragging her writhing body out into the hallway as she screamed about death and the old man.

When the door closed behind them, the teacher turned back to his class and wiped away the lone bead of sweat trickling down his forehead.

"I apologize for the disruption," he said. "Now, where were we?"

He turned back to Arthur Fink, who still held the EdPad up for him to see, his expression earnest. The space above the student's head returned to life, and after a few short moments, revealed fat, shimmering letters. The teacher paused. Smiled proudly. And read.

"Brilliant." ■

➲ NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

Jeff Samson *brews Irish stout when he's not writing science fiction, and often drinks it when he is. He lives in New Jersey with his wife and baby girl, and no cats.*

Nanostructure-enhanced atomic line emission

ARISING FROM S. Kim *et al. Nature* **453,** 757–760 (2008)

Plasmonic nanostructures offer unique possibilities for enhancing linear and nonlinear optical processes[1–6]. Recently, Kim *et al.*[7] reported nanostructure-enhanced high harmonic generation (HHG). Here, using nearly identical conditions, we demonstrate extreme-ultraviolet (EUV) emission from gas-exposed nanostructures, but come to entirely different conclusions: instead of HHG, we observe line emission of neutral and ionized gas atoms. We also discuss fundamental physical aspects limiting nanostructure-based HHG.

We conduct very similar experiments to those presented in ref. 7. Specifically, bow-tie nanostructure arrays are exposed to a noble-gas jet and illuminated with 8-fs laser pulses, and emitted radiation in the EUV is spectrally analysed (Fig. 1a). Further details and procedures are given in Methods. Figure 1b shows the raw detected spectral density of the first (solid black line) and second (solid red line) grating diffraction orders for an exemplary nanostructure (inset) and argon. We observe six main emission lines, some resolved into multiple lines in second order. All prominent features are attributed to atomic line emission of neutral and ionized argon[8,9,10]. Various optimized structures yield nearly identical spectra, whereas other gases display different transitions: Fig. 1c shows data using xenon on the same structures (Fig. 1b inset).

The presence of ionized atoms seems to support the feasibility of nanostructure-enhanced HHG. However, we have not observed any signature of HHG, even on increasing intensities beyond damage thresholds. This is a striking result, considering that the small detection solid angle in the set-up strongly favours directional emission (HHG) over incoherent line emission. There are fundamental physical reasons for the predominance of line emission in this geometry, as we discuss below. In ref. 7, using ordinary gas densities, the authors claim conversion efficiencies similar to conventional HHG. However, the much smaller number of coherently emitting dipoles, entering quadratically in the yield, suggests a huge deficit in the conversion efficiency of nanostructure-enhanced HHG. A simplified expression for the ratio of expected conversion efficiencies for nanostructure-enhanced (C_{nano}) and conventional (C_{conv}) HHG (using amplified pulses in a capillary or gas jet) at comparable local intensities is given by

$$\frac{C_{\mathrm{nano}}}{C_{\mathrm{conv}}} = \frac{R_{\mathrm{nano}}}{R_{\mathrm{conv}}}\left(\frac{N_{\mathrm{nano}}}{N_{\mathrm{conv}}}\right)^2 \frac{|F_{\mathrm{nano}}|^2}{|F_{\mathrm{conv}}|^2} \lesssim 10^{-8}$$

where $N_{\mathrm{nano}}/N_{\mathrm{conv}} \approx 10^{-8}$ and $R_{\mathrm{nano}}/R_{\mathrm{conv}} \approx 10^{5}$ are the ratios of the number of radiating atoms (at comparable density) and repetition rates in both scenarios, respectively. A typical phase matching coefficient $|F_{\mathrm{conv}}|^2 \gtrsim 10^{-3}$ is assumed for the relevant harmonics[11], while nanostructure-based HHG is assigned $|F_{\mathrm{nano}}|^2 = 1$. Such considerations may also be relevant for related studies[12]. Note that, because of a linear dependence on the dipole number for incoherent radiation, the above unfavourable conversion efficiency does not apply to atomic line emission. Thus, it is efficiently enhanced in nanostructures, as demonstrated here. In fact, a generation rate of incoherent fluorescence photons greater than $10^9\ \mathrm{s}^{-1}$ is estimated from our raw data and collection conditions.

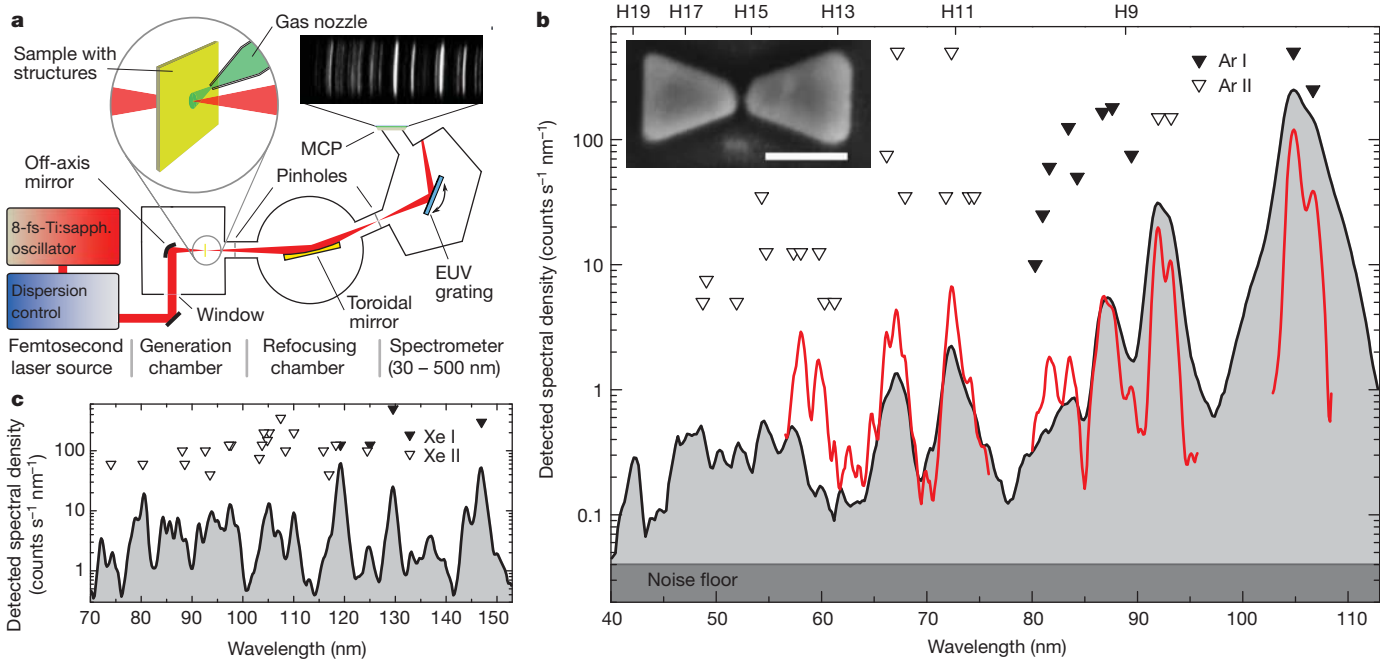


Figure 1 | Experimental set-up and results. a, Diagram of the experimental set-up. Inset, image of the phosphor screen recorded with a CCD camera. MCP, microchannel plate. This image corresponds to the xenon measurement shown in **c**. **b**, Detected spectral density (solid black line) from an array ($20 \times 20\ \mu\mathrm{m}^2$) of argon-exposed nanostructures (inset; scale bar, 200 nm) illuminated by femtosecond laser pulses. The second grating diffraction order (solid red line) provides higher resolution and efficiency at shorter wavelengths, and it is shown wherever it does not overlap with other orders. The emission corresponds to atomic line emission from neutral (Ar I; filled triangles) and singly ionized (Ar II; open triangles) argon. Vertical triangle positions indicate expected relative intensities[8–10]. Note the wavelengths expected for HHG using 800 nm light (H9–H19, upper *x*-axis). **c**, Spectrum measured using xenon and the same nanostructure as in **b** (first grating diffraction order). Triangles indicate the expected xenon lines[10,14]. Filled triangles are upshifted by a factor of 10 for better visibility.

Despite experiments with numerous high-quality samples of different dimensions (displaying efficient third harmonic generation), optimizations of gas nozzle dimensions, materials and orientations, as well as gas pressures, we have only observed atomic and ionic line emission and were able to reproduce our findings multiple times. Thus, together with the physical arguments given above, we must conclude that very efficient HHG in bow-tie nanostructures under the given conditions is highly unlikely, if not physically impossible.

We believe that our results are difficult to reconcile with the conclusions of Kim *et al.*[7], and further note several of our observations that are at variance with their results. First, in our experiments, we always observe second and higher grating diffraction orders, which is expected for broadband EUV gratings such as the ones used here and in ref. 7, where higher diffraction orders are absent. Second, the photon count rates in our experiments did not exceed several thousand per second using an imaging detector and obtaining a signal to noise ratio of $\sim 10^3$. In contrast, ref. 7 reports photon count rates above $10^8\,s^{-1}$, even exceeding the laser repetition rate, using a photon multiplier but displaying a signal to noise ratio of only $\sim 10^2$. It is very important to distinguish between actual count rates and projected generation rates, which arise from normalization by the quantum efficiency of the set-up; ref. 7 does not state which of these quantities is plotted. Generally, we believe that using conventional photon counting techniques, the nanostructure-enhanced atomic line emission we have found will be detectable in such experiments. Last, the linewidths we have measured are partially given by spectrometer resolution (below 2 nm) and are very similar to those in ref. 7 and in a related experiment with xenon[13]. Whereas harmonic linewidths are influenced by the spectral amplitude and temporal duration of the fundamental driving field, the linewidths of atomic and ionic fluorescence are governed by the spontaneous lifetime. If harmonic radiation were present, we would expect several linewidths to be broader, given the incident pulse duration and known properties of plasmonic resonances[1,3,6].

In conclusion, the line emission observed in our experiments originates from nanostructure-enhanced multiphoton and strong-field excitation and ionization, and is intrinsically incoherent. Moreover, the fundamental physical relations discussed above imply important limitations on nanostructure-enhanced HHG, which calls for alternative approaches.

METHODS

Nanostructures. Numerous arrays (area $20 \times 20\,\mu m^2$) of bow-ties are fabricated by focused ion-beam etching of smooth gold films (thermal evaporation; <1 nm r.m.s. roughness over $5 \times 5\,\mu m^2$) on EPI polished sapphire substrates. High optical (structural) nanostructure quality is confirmed using optical third harmonic generation (scanning electron/atomic force microscopy). Structural parameters are iteratively optimized for maximum emission (EUV/third harmonic), starting from nominal parameters in ref. 7. Improved nonlinear emission is found for film thicknesses, bow-tie lengths (single triangle) and gap sizes of 90 nm, 230 nm and 20 nm, respectively. For different arrays, the EUV yield depends on the field enhancement and resonance wavelength.

Experimental set-up (Fig. 1a). Optical excitation is provided by focusing dispersion-compensated 8-fs pulses from a Ti:sapphire oscillator with an off-axis parabolic mirror to incident peak intensities of 0.1–1 TW cm^{-2}. Micro-translation stages carry the samples (room temperature); a movable nozzle (stainless steel, inner diameter 100 μm) supplies a gas jet (up to 500 mbar backing pressure). The generated EUV radiation within an opening angle of $\pm 1.2^\circ$ is refocused (using a toroidal gold mirror) into a flat-field EUV spectrometer (McPherson 234, 1,200 grooves per mm). Proper alignment of the set-up for collecting possible directed radiation is ensured using the fundamental beam (zeroth grating order) and the third harmonics (267 nm) from the bare nanostructures and the substrate. EUV emission is detected with a phosphor-screen microchannel-plate assembly (Hamamatsu, uncoated). Accurate wavelength calibration is verified with plasma line emission and conventional HHG using the same set-up and nozzle.

M. Sivis[1], M. Duwe[1], B. Abel[2] & C. Ropers[1]

[1]Courant Research Center Nano-Spectroscopy and X-Ray Imaging, University of Göttingen, 37077 Göttingen, Germany.
email: cropers@gwdg.de

[2]Ostwald Institute for Physical and Theoretical Chemistry, University of Leipzig, 04103 Leipzig, Germany.

Received 5 October 2011; accepted 22 February 2012.

1. Fischer, H. & Martin, O. J. F. Engineering the optical response of plasmonic nanoantennas. *Opt. Express* **16,** 9144–9154 (2008).
2. Genov, D. A., Sarychev, A. K., Shalaev, V. M. & Wei, A. Resonant field enhancements from metal nanoparticle arrays. *Nano Lett.* **4,** 153–158 (2004).
3. Merlein, J. *et al.* Nanomechanical control of an optical antenna. *Nature Photon.* **2,** 230–233 (2008).
4. Fromm, D. P., Sundaramurthy, A., Schuck, P. J., Kino, G. & Moerner, W. E. Gap-dependent optical coupling of single "bowtie" nanoantennas resonant in the visible. *Nano Lett.* **4,** 957–961 (2004).
5. Li, K., Stockman, M. I. & Bergman, D. J. Self-similar chain of metal nanospheres as an efficient nanolens. *Phys. Rev. Lett.* **91,** 227402 (2003).
6. Hanke, T. *et al.* Efficient nonlinear light emission of single gold optical antennas driven by few-cycle near-infrared pulses. *Phys. Rev. Lett.* **103,** 257404 (2009).
7. Kim, S. *et al.* High-harmonic generation by resonant plasmon field enhancement. *Nature* **453,** 757–760 (2008).
8. Minnhagen, L. Accurately measured and calculated ground-term combinations of Ar II. *J. Opt. Soc. Am.* **61,** 1257–1262 (1971).
9. Minnhagen, L. Spectrum and the energy levels of neutral argon, Ar I. *J. Opt. Soc. Am.* **63,** 1185–1198 (1973).
10. Sansonetti, J. E. & Martin, W. C. Handbook of basic atomic spectroscopic data. *J. Phys. Chem. Ref. Data* **34,** 1582–1591; 2109–2117 (2005).
11. Li, X. F., L'Huillier, A., Ferray, M., Lompré, L. A. & Mainfray, G. Multiple-harmonic generation in rare gases at high laser intensity. *Phys. Rev. A* **39,** 5751–5761 (1989).
12. Park, I.-Y. *et al.* Plasmonic generation of ultrashort extreme-ultraviolet light pulses. *Nature Photon.* **5,** 677–681 (2011).
13. Kim, S., Park, I.-Y., Choi, J. & Kim, S.-W. in *Progress in Ultrafast Intense Laser Science* Vol. 6 (eds Yamanouchi, K., Gerber, G. & Bandrauk, A. D.) 129–144 (Springer, 2010).
14. Boyce, J. The spectra of xenon in the extreme ultraviolet. *Phys. Rev.* **49,** 730–732 (1936).

Author Contributions All authors were closely involved in this study and contributed to the ideas, realization of the experiments, data analysis and interpretation, and writing of the paper.

Competing Financial Interests Declared none.

doi:10.1038/nature10978

Kim *et al.* reply

REPLYING TO M. Sivis, M. Duwe, B. Abel & C. Ropers *Nature* **485,** http://dx.doi.org/nature10978 (2012)

Sivis *et al.* showed[1] spectral data of extreme ultraviolet (EUV) emission from gas-exposed bow-ties, claiming high predominance of atomic line emission (ALE) of neutral and ionized gas atoms in contradiction to our data[2,3] of high harmonic generation (HHG). This is not the first time the signature of ALE has been identified in conventional HHG spectral data[4]. The two distinct phenomena, ALE and HHG, are not mutually exclusive but coexistent when gaseous atoms are illuminated by strong-field laser pulses.

The feasibility of nanostructure-enhanced HHG has been proved already by several theoretical studies conducted independently[5–8]. However, no experimental demonstration has yet been reported except for our HHG data[2,3] from bow-ties. The main reason may be inferred from the durability problem we encountered with bow-ties patterned with Au on a sapphire substrate. The thin-film bow-ties began to degrade, not abruptly, but starting gradually then continuing at a fast rate, soon after being exposed to the driving laser set to 0.1 TW cm^{-2} intensity. Being continuously deprived of geometrical accuracy by thermal damage coupled with optical breakdown, even new bow-ties gave out detectable HHG signals during a short lifetime, which was often missed.

We do not agree with Sivis *et al.*[1] that $N_{nano}/N_{conv} \approx 10^{-8}$. Our calculation reveals that the ratio reaches 10^{-6}; we calculate the number of gas atoms to be $\sim 8 \times 10^4$ with a total interaction volume of $60\,nm \times 50\,nm \times 50\,nm \times 150$ bow-ties at 115 torr pressure. This brings the conversion efficiency of nano-HHG for the seventh harmonic (H7) to be $\sim 10^{-4}$ times that of conventional HHG. Given that the conversion efficiency of H7 is $\sim 10^{-5}$ in conventional HHG[9], the efficiency value of 10^{-9} for the harmonic shown in our data is realistic. No consideration was given to the possible enhancement of harmonic yield attributable to inhomogeneous distribution of plasmonic field intensity[8].

The photon-count of our data[2,3] was estimated by measuring the output current of the photomultiplier tube (PMT) used to scan the raw spectral data, and at the same time taking into account all the individual functional efficiencies of the hardware components involved in our experiment. This projected estimation of photon-count led to a noise floor of $\sim 10^6$ photons s^{-1}, which appeared rather high due to electrical noise in the PMT current measurement.

The linewidth became narrowed during post-processing, both for deconvolution of the slit size placed before the PMT and also for data averaging through repetitive measurements to reduce electric noise. Long plasmonic field decay might have caused the linewidth narrowing[7], which is however not proved yet. No attention was paid to the second or higher-order diffraction signals because they were buried below the noise floor in our measurement; the grating efficiency for the second order diffraction was one order of magnitude less than that for the first order. Besides, the peak amplitudes of higher harmonics were also about one order of magnitude less than that of H7.

Further work continued after our Letter[2] brought us the conclusion that bow-ties are not an ideal tool for experimental demonstration of nano-HHG. Instead, we found that three-dimensional waveguides[10] hollowed out in an ellipsoidal funnel shape on a bulk metal substrate are a good alternative. The funnel waveguide permits stable, consistent generation of higher harmonics up to the 43rd order using xenon gas with improved immunity to optical and thermal damage. Investigation is underway to verify the coherence of the EUV radiation from the funnel waveguide.

Seungchul Kim[1], Jonghan Jin[1], Young-Jin Kim[1], In-Yong Park[1], Yunseok Kim[1] & Seung-Woo Kim[1]

[1]Billionth Uncertainty Precision Engineering Group, KAIST, Daedeok Science Town, Daejeon 305-701, South Korea.
email: swk@kaist.ac.kr

1. Sivis, M., Duwe, M., Abel, B. & Ropers, C. Nanostructure-enhanced atomic line emission. *Nature* **485,** http://dx.doi.org/nature10978 (2012).
2. Kim, S. *et al.* High harmonic generation by resonant plasmon field enhancement. *Nature* **453,** 757–760 (2008).
3. Kim, S. *et al.* in *Progress in Ultrafast Intense Laser Science* Vol. 6 (eds Yamanouchi, K., Gerber, G. & Bandrauk, A. D.) 129–144 (Springer, 2010).
4. Gohle, C. *et al.* A frequency comb in the extreme ultraviolet. *Nature* **436,** 234–237 (2005).
5. Husakou, A. *et al.* Theory of plasmon-enhanced high-order harmonic generation in the vicinity of metal nanostructures in noble gases. *Phys. Rev. A* **83,** 043839 (2011).
6. Husakou, A. *et al.* Polarization gating and circularly-polarized high harmonic generation using plasmonic enhancement in metal nanostructures. *Opt. Express* **19,** 25346 (2011).
7. Stebbings, S. L. *et al.* Generation of isolated attosecond extreme ultraviolet pulses employing nanoplasmonic field enhancement: optimization of coupled ellipsoids. *N. J. Phys.* **13,** 073010 (2011).
8. Yavuz, I. Generation of a broadband XUV continuum in high-order-harmonic generation by spatially inhomogeneous fields. *Phys. Rev. A* **85,** 013416 (2012).
9. Popmintchev, T. *et al.* Phase matching of high harmonic generation in the soft and hard X-ray regions of the spectrum. *Proc. Natl Acad. Sci. USA* **106,** 10516–10521 (2009).
10. Park, I. Y. *et al.* Plasmonic generation of ultrashort extreme-ultraviolet light pulses. *Nature Photon.* **5,** 677–681 (2011).

Author Contributions This Reply was written by S.-W.K. and S.K. on behalf of all the authors.

doi:10.1038/nature10979

Nanostructure-enhanced atomic line emission

ARISING FROM S. Kim *et al. Nature* **453,** 757–760 (2008)

Plasmonic nanostructures offer unique possibilities for enhancing linear and nonlinear optical processes[1–6]. Recently, Kim *et al.*[7] reported nanostructure-enhanced high harmonic generation (HHG). Here, using nearly identical conditions, we demonstrate extreme-ultraviolet (EUV) emission from gas-exposed nanostructures, but come to entirely different conclusions: instead of HHG, we observe line emission of neutral and ionized gas atoms. We also discuss fundamental physical aspects limiting nanostructure-based HHG.

We conduct very similar experiments to those presented in ref. 7. Specifically, bow-tie nanostructure arrays are exposed to a noble-gas jet and illuminated with 8-fs laser pulses, and emitted radiation in the EUV is spectrally analysed (Fig. 1a). Further details and procedures are given in Methods. Figure 1b shows the raw detected spectral density of the first (solid black line) and second (solid red line) grating diffraction orders for an exemplary nanostructure (inset) and argon. We observe six main emission lines, some resolved into multiple lines in second order. All prominent features are attributed to atomic line emission of neutral and ionized argon[8,9,10]. Various optimized structures yield nearly identical spectra, whereas other gases display different transitions: Fig. 1c shows data using xenon on the same structures (Fig. 1b inset).

The presence of ionized atoms seems to support the feasibility of nanostructure-enhanced HHG. However, we have not observed any signature of HHG, even on increasing intensities beyond damage thresholds. This is a striking result, considering that the small detection solid angle in the set-up strongly favours directional emission (HHG) over incoherent line emission. There are fundamental physical reasons for the predominance of line emission in this geometry, as we discuss below. In ref. 7, using ordinary gas densities, the authors claim conversion efficiencies similar to conventional HHG. However, the much smaller number of coherently emitting dipoles, entering quadratically in the yield, suggests a huge deficit in the conversion efficiency of nanostructure-enhanced HHG. A simplified expression for the ratio of expected conversion efficiencies for nanostructure-enhanced (C_{nano}) and conventional (C_{conv}) HHG (using amplified pulses in a capillary or gas jet) at comparable local intensities is given by

$$\frac{C_{\mathrm{nano}}}{C_{\mathrm{conv}}} = \frac{R_{\mathrm{nano}}}{R_{\mathrm{conv}}}\left(\frac{N_{\mathrm{nano}}}{N_{\mathrm{conv}}}\right)^2 \frac{|F_{\mathrm{nano}}|^2}{|F_{\mathrm{conv}}|^2} \lesssim 10^{-8}$$

where $N_{\mathrm{nano}}/N_{\mathrm{conv}} \approx 10^{-8}$ and $R_{\mathrm{nano}}/R_{\mathrm{conv}} \approx 10^{5}$ are the ratios of the number of radiating atoms (at comparable density) and repetition rates in both scenarios, respectively. A typical phase matching coefficient $|F_{\mathrm{conv}}|^2 \gtrsim 10^{-3}$ is assumed for the relevant harmonics[11], while nanostructure-based HHG is assigned $|F_{\mathrm{nano}}|^2 = 1$. Such considerations may also be relevant for related studies[12]. Note that, because of a linear dependence on the dipole number for incoherent radiation, the above unfavourable conversion efficiency does not apply to atomic line emission. Thus, it is efficiently enhanced in nanostructures, as demonstrated here. In fact, a generation rate of incoherent fluorescence photons greater than $10^9\ \mathrm{s}^{-1}$ is estimated from our raw data and collection conditions.

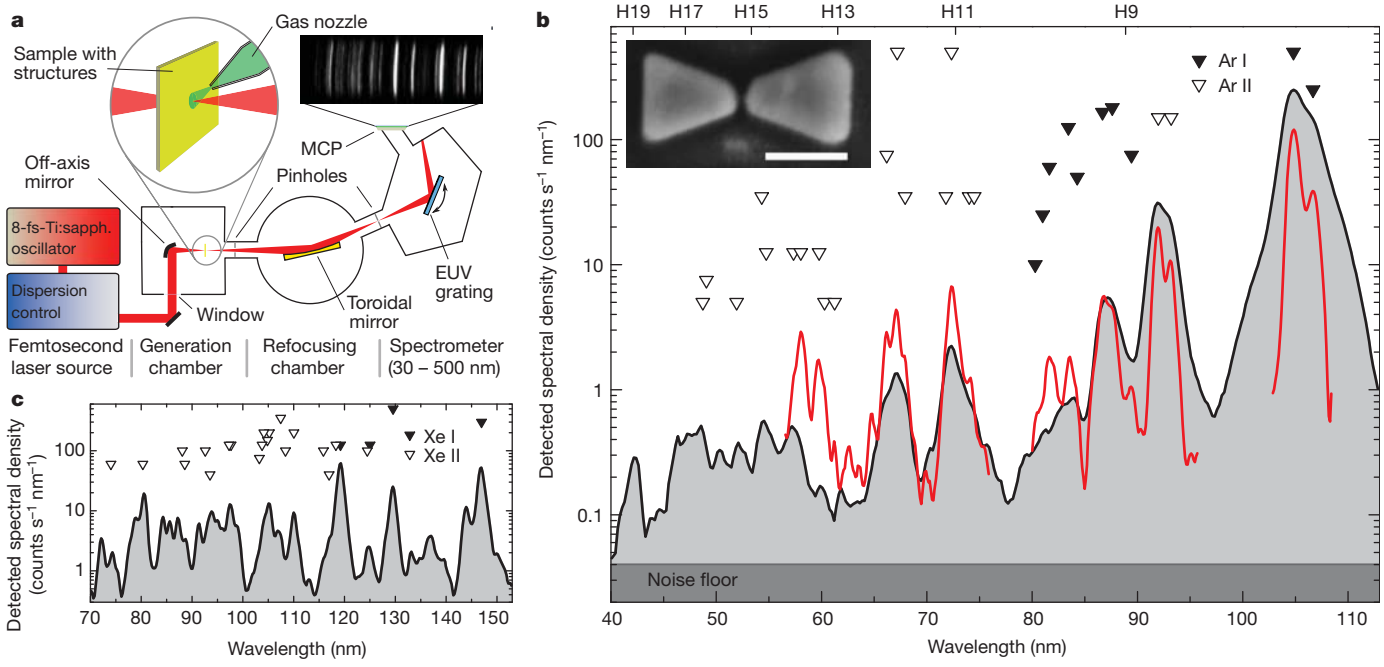


Figure 1 | Experimental set-up and results. a, Diagram of the experimental set-up. Inset, image of the phosphor screen recorded with a CCD camera. MCP, microchannel plate. This image corresponds to the xenon measurement shown in **c**. **b**, Detected spectral density (solid black line) from an array ($20 \times 20\ \mu\mathrm{m}^2$) of argon-exposed nanostructures (inset; scale bar, 200 nm) illuminated by femtosecond laser pulses. The second grating diffraction order (solid red line) provides higher resolution and efficiency at shorter wavelengths, and it is shown wherever it does not overlap with other orders. The emission corresponds to atomic line emission from neutral (Ar I; filled triangles) and singly ionized (Ar II; open triangles) argon. Vertical triangle positions indicate expected relative intensities[8–10]. Note the wavelengths expected for HHG using 800 nm light (H9–H19, upper *x*-axis). **c**, Spectrum measured using xenon and the same nanostructure as in **b** (first grating diffraction order). Triangles indicate the expected xenon lines[10,14]. Filled triangles are upshifted by a factor of 10 for better visibility.

Despite experiments with numerous high-quality samples of different dimensions (displaying efficient third harmonic generation), optimizations of gas nozzle dimensions, materials and orientations, as well as gas pressures, we have only observed atomic and ionic line emission and were able to reproduce our findings multiple times. Thus, together with the physical arguments given above, we must conclude that very efficient HHG in bow-tie nanostructures under the given conditions is highly unlikely, if not physically impossible.

We believe that our results are difficult to reconcile with the conclusions of Kim *et al.*[7], and further note several of our observations that are at variance with their results. First, in our experiments, we always observe second and higher grating diffraction orders, which is expected for broadband EUV gratings such as the ones used here and in ref. 7, where higher diffraction orders are absent. Second, the photon count rates in our experiments did not exceed several thousand per second using an imaging detector and obtaining a signal to noise ratio of $\sim 10^3$. In contrast, ref. 7 reports photon count rates above $10^8\,s^{-1}$, even exceeding the laser repetition rate, using a photon multiplier but displaying a signal to noise ratio of only $\sim 10^2$. It is very important to distinguish between actual count rates and projected generation rates, which arise from normalization by the quantum efficiency of the set-up; ref. 7 does not state which of these quantities is plotted. Generally, we believe that using conventional photon counting techniques, the nanostructure-enhanced atomic line emission we have found will be detectable in such experiments. Last, the linewidths we have measured are partially given by spectrometer resolution (below 2 nm) and are very similar to those in ref. 7 and in a related experiment with xenon[13]. Whereas harmonic linewidths are influenced by the spectral amplitude and temporal duration of the fundamental driving field, the linewidths of atomic and ionic fluorescence are governed by the spontaneous lifetime. If harmonic radiation were present, we would expect several linewidths to be broader, given the incident pulse duration and known properties of plasmonic resonances[1,3,6].

In conclusion, the line emission observed in our experiments originates from nanostructure-enhanced multiphoton and strong-field excitation and ionization, and is intrinsically incoherent. Moreover, the fundamental physical relations discussed above imply important limitations on nanostructure-enhanced HHG, which calls for alternative approaches.

METHODS

Nanostructures. Numerous arrays (area $20 \times 20\,\mu m^2$) of bow-ties are fabricated by focused ion-beam etching of smooth gold films (thermal evaporation; <1 nm r.m.s. roughness over $5 \times 5\,\mu m^2$) on EPI polished sapphire substrates. High optical (structural) nanostructure quality is confirmed using optical third harmonic generation (scanning electron/atomic force microscopy). Structural parameters are iteratively optimized for maximum emission (EUV/third harmonic), starting from nominal parameters in ref. 7. Improved nonlinear emission is found for film thicknesses, bow-tie lengths (single triangle) and gap sizes of 90 nm, 230 nm and 20 nm, respectively. For different arrays, the EUV yield depends on the field enhancement and resonance wavelength.

Experimental set-up (Fig. 1a). Optical excitation is provided by focusing dispersion-compensated 8-fs pulses from a Ti:sapphire oscillator with an off-axis parabolic mirror to incident peak intensities of 0.1–1 TW cm^{-2}. Micro-translation stages carry the samples (room temperature); a movable nozzle (stainless steel, inner diameter 100 μm) supplies a gas jet (up to 500 mbar backing pressure). The generated EUV radiation within an opening angle of $\pm 1.2^\circ$ is refocused (using a toroidal gold mirror) into a flat-field EUV spectrometer (McPherson 234, 1,200 grooves per mm). Proper alignment of the set-up for collecting possible directed radiation is ensured using the fundamental beam (zeroth grating order) and the third harmonics (267 nm) from the bare nanostructures and the substrate. EUV emission is detected with a phosphor-screen microchannel-plate assembly (Hamamatsu, uncoated). Accurate wavelength calibration is verified with plasma line emission and conventional HHG using the same set-up and nozzle.

M. Sivis[1], M. Duwe[1], B. Abel[2] & C. Ropers[1]

[1]Courant Research Center Nano-Spectroscopy and X-Ray Imaging, University of Göttingen, 37077 Göttingen, Germany.
email: cropers@gwdg.de

[2]Ostwald Institute for Physical and Theoretical Chemistry, University of Leipzig, 04103 Leipzig, Germany.

Received 5 October 2011; accepted 22 February 2012.

1. Fischer, H. & Martin, O. J. F. Engineering the optical response of plasmonic nanoantennas. *Opt. Express* **16,** 9144–9154 (2008).
2. Genov, D. A., Sarychev, A. K., Shalaev, V. M. & Wei, A. Resonant field enhancements from metal nanoparticle arrays. *Nano Lett.* **4,** 153–158 (2004).
3. Merlein, J. *et al.* Nanomechanical control of an optical antenna. *Nature Photon.* **2,** 230–233 (2008).
4. Fromm, D. P., Sundaramurthy, A., Schuck, P. J., Kino, G. & Moerner, W. E. Gap-dependent optical coupling of single "bowtie" nanoantennas resonant in the visible. *Nano Lett.* **4,** 957–961 (2004).
5. Li, K., Stockman, M. I. & Bergman, D. J. Self-similar chain of metal nanospheres as an efficient nanolens. *Phys. Rev. Lett.* **91,** 227402 (2003).
6. Hanke, T. *et al.* Efficient nonlinear light emission of single gold optical antennas driven by few-cycle near-infrared pulses. *Phys. Rev. Lett.* **103,** 257404 (2009).
7. Kim, S. *et al.* High-harmonic generation by resonant plasmon field enhancement. *Nature* **453,** 757–760 (2008).
8. Minnhagen, L. Accurately measured and calculated ground-term combinations of Ar II. *J. Opt. Soc. Am.* **61,** 1257–1262 (1971).
9. Minnhagen, L. Spectrum and the energy levels of neutral argon, Ar I. *J. Opt. Soc. Am.* **63,** 1185–1198 (1973).
10. Sansonetti, J. E. & Martin, W. C. Handbook of basic atomic spectroscopic data. *J. Phys. Chem. Ref. Data* **34,** 1582–1591; 2109–2117 (2005).
11. Li, X. F., L'Huillier, A., Ferray, M., Lompré, L. A. & Mainfray, G. Multiple-harmonic generation in rare gases at high laser intensity. *Phys. Rev. A* **39,** 5751–5761 (1989).
12. Park, I.-Y. *et al.* Plasmonic generation of ultrashort extreme-ultraviolet light pulses. *Nature Photon.* **5,** 677–681 (2011).
13. Kim, S., Park, I.-Y., Choi, J. & Kim, S.-W. in *Progress in Ultrafast Intense Laser Science* Vol. 6 (eds Yamanouchi, K., Gerber, G. & Bandrauk, A. D.) 129–144 (Springer, 2010).
14. Boyce, J. The spectra of xenon in the extreme ultraviolet. *Phys. Rev.* **49,** 730–732 (1936).

Author Contributions All authors were closely involved in this study and contributed to the ideas, realization of the experiments, data analysis and interpretation, and writing of the paper.

Competing Financial Interests Declared none.

doi:10.1038/nature10978

Kim *et al.* reply

REPLYING TO M. Sivis, M. Duwe, B. Abel & C. Ropers *Nature* **485,** http://dx.doi.org/nature10978 (2012)

Sivis *et al.* showed[1] spectral data of extreme ultraviolet (EUV) emission from gas-exposed bow-ties, claiming high predominance of atomic line emission (ALE) of neutral and ionized gas atoms in contradiction to our data[2,3] of high harmonic generation (HHG). This is not the first time the signature of ALE has been identified in conventional HHG spectral data[4]. The two distinct phenomena, ALE and HHG, are not mutually exclusive but coexistent when gaseous atoms are illuminated by strong-field laser pulses.

The feasibility of nanostructure-enhanced HHG has been proved already by several theoretical studies conducted independently[5–8]. However, no experimental demonstration has yet been reported except for our HHG data[2,3] from bow-ties. The main reason may be inferred from the durability problem we encountered with bow-ties patterned with Au on a sapphire substrate. The thin-film bow-ties began to degrade, not abruptly, but starting gradually then continuing at a fast rate, soon after being exposed to the driving laser set to 0.1 TW cm^{-2} intensity. Being continuously deprived of geometrical accuracy by thermal damage coupled with optical breakdown, even new bow-ties gave out detectable HHG signals during a short lifetime, which was often missed.

We do not agree with Sivis *et al.*[1] that $N_{nano}/N_{conv} \approx 10^{-8}$. Our calculation reveals that the ratio reaches 10^{-6}; we calculate the number of gas atoms to be $\sim 8 \times 10^4$ with a total interaction volume of $60\,nm \times 50\,nm \times 50\,nm \times 150$ bow-ties at 115 torr pressure. This brings the conversion efficiency of nano-HHG for the seventh harmonic (H7) to be $\sim 10^{-4}$ times that of conventional HHG. Given that the conversion efficiency of H7 is $\sim 10^{-5}$ in conventional HHG[9], the efficiency value of 10^{-9} for the harmonic shown in our data is realistic. No consideration was given to the possible enhancement of harmonic yield attributable to inhomogeneous distribution of plasmonic field intensity[8].

The photon-count of our data[2,3] was estimated by measuring the output current of the photomultiplier tube (PMT) used to scan the raw spectral data, and at the same time taking into account all the individual functional efficiencies of the hardware components involved in our experiment. This projected estimation of photon-count led to a noise floor of $\sim 10^6$ photons s^{-1}, which appeared rather high due to electrical noise in the PMT current measurement.

The linewidth became narrowed during post-processing, both for deconvolution of the slit size placed before the PMT and also for data averaging through repetitive measurements to reduce electric noise. Long plasmonic field decay might have caused the linewidth narrowing[7], which is however not proved yet. No attention was paid to the second or higher-order diffraction signals because they were buried below the noise floor in our measurement; the grating efficiency for the second order diffraction was one order of magnitude less than that for the first order. Besides, the peak amplitudes of higher harmonics were also about one order of magnitude less than that of H7.

Further work continued after our Letter[2] brought us the conclusion that bow-ties are not an ideal tool for experimental demonstration of nano-HHG. Instead, we found that three-dimensional waveguides[10] hollowed out in an ellipsoidal funnel shape on a bulk metal substrate are a good alternative. The funnel waveguide permits stable, consistent generation of higher harmonics up to the 43rd order using xenon gas with improved immunity to optical and thermal damage. Investigation is underway to verify the coherence of the EUV radiation from the funnel waveguide.

Seungchul Kim[1], Jonghan Jin[1], Young-Jin Kim[1], In-Yong Park[1], Yunseok Kim[1] & Seung-Woo Kim[1]

[1]Billionth Uncertainty Precision Engineering Group, KAIST, Daedeok Science Town, Daejeon 305-701, South Korea.
email: swk@kaist.ac.kr

1. Sivis, M., Duwe, M., Abel, B. & Ropers, C. Nanostructure-enhanced atomic line emission. *Nature* **485,** http://dx.doi.org/nature10978 (2012).
2. Kim, S. *et al.* High harmonic generation by resonant plasmon field enhancement. *Nature* **453,** 757–760 (2008).
3. Kim, S. *et al.* in *Progress in Ultrafast Intense Laser Science* Vol. 6 (eds Yamanouchi, K., Gerber, G. & Bandrauk, A. D.) 129–144 (Springer, 2010).
4. Gohle, C. *et al.* A frequency comb in the extreme ultraviolet. *Nature* **436,** 234–237 (2005).
5. Husakou, A. *et al.* Theory of plasmon-enhanced high-order harmonic generation in the vicinity of metal nanostructures in noble gases. *Phys. Rev. A* **83,** 043839 (2011).
6. Husakou, A. *et al.* Polarization gating and circularly-polarized high harmonic generation using plasmonic enhancement in metal nanostructures. *Opt. Express* **19,** 25346 (2011).
7. Stebbings, S. L. *et al.* Generation of isolated attosecond extreme ultraviolet pulses employing nanoplasmonic field enhancement: optimization of coupled ellipsoids. *N. J. Phys.* **13,** 073010 (2011).
8. Yavuz, I. Generation of a broadband XUV continuum in high-order-harmonic generation by spatially inhomogeneous fields. *Phys. Rev. A* **85,** 013416 (2012).
9. Popmintchev, T. *et al.* Phase matching of high harmonic generation in the soft and hard X-ray regions of the spectrum. *Proc. Natl Acad. Sci. USA* **106,** 10516–10521 (2009).
10. Park, I. Y. *et al.* Plasmonic generation of ultrashort extreme-ultraviolet light pulses. *Nature Photon.* **5,** 677–681 (2011).

Author Contributions This Reply was written by S.-W.K. and S.K. on behalf of all the authors.

doi:10.1038/nature10979

FORUM: Agriculture

Comparing apples with oranges

A meta-analysis of agricultural systems shows that organic yields are mostly lower than those from conventional farming, but that organic crops perform well in some contexts. Agricultural scientists discuss whether the conclusions of the study should change farming practices and management. SEE LETTER P.229

THE PAPER IN BRIEF

- A growing human population poses challenges to agricultural sustainability and food security.
- Organic farming is deemed less environmentally damaging than non-organic systems, but it may require more land to produce the same amount of food.
- Seufert *et al.*[1] (page 229) did a categorized analysis of existing data to compare the efficiency of the two agricultural approaches.
- The authors find that, although organic yields are lower on average, they are almost equivalent to conventional yields for some crop types and when good organic management practices are used.

The fruits of organic farming

JOHN P. REGANOLD

Yield differences between organic and conventional farming systems are a topic of intensive debate, and numerous studies have compared crop yields. Yet few studies have synthesized this information on a global scale. In a meta-analysis, Seufert *et al.*[1] show, from 316 yield comparisons in 66 studies, that organic farming systems in developed countries produce yields that are 20% lower than their conventional counterparts. This discrepancy rises to 25% when data from developed and developing countries are combined. However, the authors also found that for certain crops (Fig. 1), growing conditions and management practices, organic yields nearly match those from conventional systems. These findings underscore the potential for organic farming to have an increasing role in a sustainable food supply.

In the first extensive review of organic versus conventional yield data, conducted in 1990, Stanhill[2] found organic yields to be 9% lower than conventional yields in developed countries. A subsequent study by Badgley *et al.*[3] found this difference to be 8%. In another recent meta-analysis of 362 yield comparisons, de Ponti and colleagues[4] found organic yields to be 21% lower in developed countries and 20% lower globally. In addition, they found that the best-yielding organically grown crops are rice (6% lower yield than conventional), soya beans (8% lower), corn (11% lower) and grass–clover (11% lower). In comparison, the highest-yielding organic crops identified by Seufert and colleagues were organic fruits (3% lower yield than conventional), rain-fed legumes such as soya beans (5% lower) and oilseed crops (11% lower).

One likely reason for greater average-yield differences in the Seufert *et al.*[1] and de Ponti *et al.*[4] meta-analyses, compared with the earlier studies[2,3], is their more restrictive selection criteria. For example, Seufert and colleagues excluded 268 possible yield comparisons simply because the studies failed to report sample size or estimates of standard error. As the authors admit, their criteria also biased the analysis of yields in developing countries by using atypically higher conventional yields — in 58 of their 67 comparisons in this category, conventional yields were more than 50% higher than average yields from the respective regions.

Nevertheless, Seufert and colleagues reveal remarkable findings when the systems are further stratified according to different categories, such as crop type, level of management and the stage of crop growth — an example of meta-analysis being a great tool for identifying broad patterns not immediately visible in primary field research[5]. Although this analysis technique must also be treated with caution (because no single farming system or practice works best in every location), both the Seufert *et al.* and de Ponti *et al.* studies bolster the argument that adoption of organic agriculture under conditions in which it performs best might close the yield gap between organic and conventional systems.

If we want to feed a growing world population, producing adequate crop yields is vital. But, as described in a report[6] by the US National Research Council (NRC), sufficient productivity is only one of four main goals that must be met for agriculture to be sustainable. The other three are enhancing the natural-resource base and environment, making farming financially viable, and contributing to the well-being of farmers and their communities. Conventional farming systems have provided increasing supplies of food and other products, but often at the expense of the other three sustainability goals. The NRC report[6] identifies organic methods as one of several innovative systems that better integrate production, environmental and socio-economic objectives. Other such systems include agroforestry, hybrid organic–conventional agriculture, conservation agriculture, grass-fed livestock production and mixed crop–livestock systems.

No one of these systems alone will produce enough food to feed the planet. Rather, a blend of farming approaches is needed for future global food and ecosystem security. Organic farming provides multiple sustainability benefits, and Seufert and colleagues' findings indicate that it can play a part in feeding the world. Yet just under 1% of agricultural land worldwide is now managed organically[7]. This percentage should be much larger in the future.

John P. Reganold *is in the Department of Crop and Soil Sciences, Washington State University, Pullman, Washington 99164, USA.*
e-mail: reganold@wsu.edu

Getting back to the field

ACHIM DOBERMANN

Seufert and colleagues[1] have added another meta-analysis to the popular debate on whether organic agriculture systems can feed the world, which joins a similar recent analysis by de Ponti *et al.*[4]. Both studies report similar results: that yields of well-managed organically grown crops average about 75–80% of the crop yield under conventional management, and that


L. VEIGA/GETTY IMAGES

Figure 1 | Crunch time for agriculture. Seufert and colleagues' meta-analysis[1] shows that for some organic crops, such as apples, farming yields almost match those from conventional agriculture.

the size of the yield gap is highly contextual. The more rigorous selection criteria and analysis methods used in these two studies make them a substantial improvement on previous studies that suggested only slightly lower organic yields or yields that even exceeded those obtained with conventional farming. The analysis methods used to derive such estimates, particularly those used by Badgley *et al.*[3], were, in my and others' opinion[8], questionable, owing to their reliance on yield ratios that in many cases represented large differences in crop management, particularly in nutrient inputs.

Despite the valuable contribution of the two new studies, the results are hardly surprising. Any experienced agronomist or farmer knows that achieving a high crop yield requires a well-adapted plant variety, sufficient sunshine, water and nutrients, and good soil and crop care. These prerequisites do not differ between conventional and organic agriculture.

It is time to accept that various types of agriculture can have a place in feeding the world, depending on the availability of land, the degree of self-reliance of agricultural systems in terms of critical inputs to value chains (such as nutrients and other resources), the scale of food production, and the desired and feasible trade in agricultural goods[9]. But we also need to leave vague, outdated concepts of sustainability behind, because the real picture is much more complex than it seems. Organic or low-external-input agriculture is not always sustainable[10]. There are also many conventional agricultural systems that are highly productive, resource-efficient and sustainable[11] — and some have been so for a long time. Instead of doing further meta-analyses to attempt to determine the optimal combination of agricultural systems, scientists should return to their fields and laboratories, and concentrate their efforts on increasing the performance of both conventional and organic agriculture.

What should scientists study? As de Ponti *et al.* point out, one issue is the scaling-up of organic agriculture. Side-by-side comparisons at the field or plot scale have shown that ensuring a sustainable, cost-effective supply of plant nutrients is a key constraint in organic systems, irrespective of whether the materials providing the nutrients are organic or mineral. Therein lies the biggest challenge for larger-scale organic agriculture. The most relevant parameter for food security and for preserving natural ecosystems is food output per unit area-time — which should ideally be optimized on existing agricultural land. But land, time, labour, money and transport are required to produce and distribute nutrients from organic sources. Where would the extra land to grow the extra nutrients be found? Comparative studies are needed to assess what scale of organic agriculture might be feasible from a nutrient capture and transfer point of view, and where this could be done.

We also need more evidence that organic agriculture systems can be designed so that they do not require premium prices or government subsidies to remain economically viable. If that cannot be shown, how will we progress in the fight to combat poverty, hunger and malnutrition in developing nations?

Yield and input-efficiency gaps exist in both organic and conventional agriculture. Closing these gaps and meeting high profitability, environmental, sustainability and social standards are not mutually exclusive goals, but the value chain (from seed to table) that should be implemented will depend on local conditions. Fine-tuning these requirements requires a more accurate understanding of crop yield potential, yield gaps, resource efficiencies, environmental impact and sustainability in quantitative terms than we have currently — to provide us with the precise agricultural technologies needed to reach higher performance and sustainability standards. Comparing one system with another in relative terms will not enhance our understanding of the requirements for a better yield, but well-designed experimental research at scales relevant to the production level may. ■

Achim Dobermann *is at the International Rice Research Institute (IRRI), Metro Manila 1301, Philippines.*
e-mail: a.dobermann@irri.org

1. Seufert, V., Ramankutty, N. & Foley, J. A. *Nature* **485,** 229–232 (2012).
2. Stanhill, G. *Agr. Ecosyst. Environ.* **30,** 1–26 (1990).
3. Badgley, C. *et al. Renew. Agr. Food Syst.* **22,** 86–108 (2007).
4. de Ponti, T., Rijk, B. & van Ittersum, M. K. *Agr. Syst.* **108,** 1–9 (2012).
5. Arnqvist, G. & Wooster, D. *Trends Ecol. Evol.* **10,** 236–240 (1995).
6. NRC *Toward Sustainable Agricultural Systems in the 21st Century* (National Academies Press, 2010).
7. Willer, H. in *The World of Organic Agriculture: Statistics and Emerging Trends 2011* (eds Willer, H. & Kilcher, L.) 26–32 (IFOAM & FiBL, 2011).
8. Connor, D. J. *Field Crops Res.* **106,** 187–190 (2008).
9. Rigby, D. & Caceres, D. *Agric. Syst.* **68,** 21–40 (2001).
10. Leifeld, J. *Agric. Ecosyst. Environ.* **150,** 121–122 (2012).
11. Grassini, P. & Cassman, K. G. *Proc. Natl Acad. Sci. USA* **109,** 1074–1079 (2012).

MATERIALS SCIENCE

Cracks tamed

Crack propagation in materials is rarely welcome. But carefully engineered cracks produced during the deposition of a film on silicon can be used to efficiently create pre-designed patterns of nanometre-scale channels. SEE LETTER P.221

ANTONIO J. PONS

The potential mechanical energy stored at the interface between a film and an underlying crystal substrate[1] can, in some cases, be released in the form of a crack that propagates through the laminate material (Fig. 1). Fissures produced in this way usually spread freely. But on page 221 of this issue, Nam *et al.*[2] present a technique that fully controls such fracture progression. As a result, the authors were able to fabricate microscopic patterns of cracks, thereby unveiling a promising alternative to other high-resolution approaches to making patterns of channels on surfaces for applications in fields such as

electronics and microfluidics.

Many techniques in daily use for producing high-resolution nanopatterns are based on lithography and etching. Lithography is a process whereby light[3], electrons[4] or ions[5] are used to create high-resolution patterns of holes, channels or more complex structures in 'resist' materials on top of a wafer. During the subsequent etching process, those layers of the wafer that are not protected by the resist are selectively removed, imprinting the patterns onto the wafer.

Although the usefulness of these techniques has been proved, the methods are frequently complex and costly, and have a number of drawbacks. For instance, lithography is often time-consuming, which restricts its throughput, and its resolution can be limited by factors such as diffraction or electron back-scattering. Etching processes, meanwhile, can remove non-target parts of the substrate material. New methods for making high-resolution nanopatterns cheaply, and with high throughput, are therefore always welcomed by industry.

In their approach, Nam *et al.*[2] take advantage of a known phenomenon: when a thin layer of silicon nitride is deposited on a silicon substrate, uncontrolled cracks can appear in the film and/or the substrate. By imprinting several features onto the substrate, the authors control the appearance — the direction, size and morphology — of such cracks. In this way, desired patterns can be created simply during the deposition process, avoiding the problems associated with other techniques.

To initiate crack formation, Nam *et al.* tailored micro-notches in the silicon to have tip angles that concentrated stresses to levels above the failure threshold of the system. Once a crack had initiated, the authors observed that it formed one of three morphologies: straight, oscillatory or stitch-like (see Fig. 1b of the paper[2]). The precise shapes were determined by several factors, such as the conditions used to deposit the silicon nitride, the orientation of layers of atoms in the substrate's crystal lattice, and the presence or absence of a silicon dioxide interlayer between the silicon nitride film and the silicon substrate. The fabricated channels had widths ranging from 10 nanometres for straight cracks to about 120 nanometres for oscillatory ones.

The researchers also generated step-like profiles — terraced edges — in the silicon substrate to stop crack propagation, and used them to protect areas of the crystal from crack penetration. In the absence of these stopper structures, crack propagation is limited only by the size of the silicon wafer, in principle allowing channels to be produced with lengths in the decimetre range. Channels that have such high aspect ratios (the ratio of length to width) are difficult to make using standard lithographic techniques. Another advantage of Nam and colleagues' technique is that it takes only a matter of hours to produce a pattern, irrespective of the wafer size; other techniques, such as electron-beam lithography, would require days or weeks for wafer sizes of a few square centimetres.

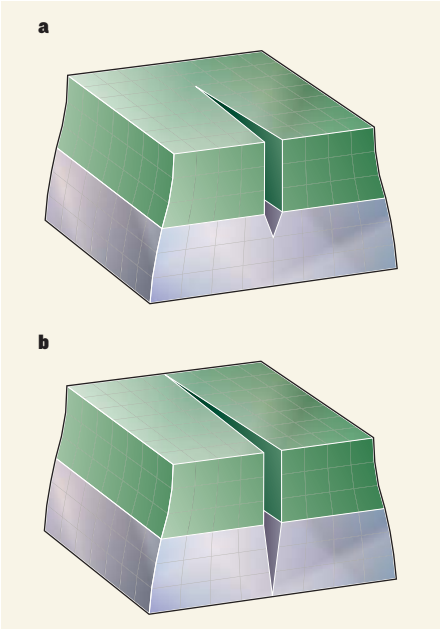


Figure 1 | Cracks produced at crystal interfaces. a, When layers of materials that have non-matching crystal lattices are put together, stress builds up at the interface. In many cases, the resulting deformation produces a gap that is maintained without crack propagation. Each small 'block' represents an atom or a molecule. **b**, On other occasions, the potential energy generated by the deformation is released, breaking atomic bonds and creating a crack that propagates in both layers.

Given that the anisotropic — direction-dependent — arrangement of atoms in crystal lattices determines the path of crack propagation during failure processes[6], Nam *et al.* were also able to drive the direction of crack propagation in their system in a way reminiscent of the refraction of light (see Fig. 3a of the paper[2]). They did this by modifying the anisotropy of the substrate, introducing regions that contained a layer of silicon dioxide between the film and the substrate, and other regions that did not. They observed that the absence of an interlayer increased the penetration of cracks into the substrate, amplifying the influence of the orientation of atomic layers in the substrate on the stress distribution that drives the direction of crack propagation. In other words, when a crack progressed through the different regions, its direction of propagation changed in response to the substrate's crystallographic anisotropy.

By developing precise ways to manipulate crack dynamics, Nam *et al.* have provided a set of tools for imprinting artificial microscopic patterns. The potential impact of these results on different technologies is high. For instance, such a simple but accurate technique is fully compatible with standard silicon-based integrated circuits, and could potentially be used in the semiconductor industry as part of the construction process to create chips. Furthermore, the channels produced with this technique might find use as templates to make nanowires with predefined geometries. Patterns of cracks could also be used in conjunction with nanoimprint lithography[7] to produce sophisticated moulds, embossed into a thin layer of liquid silicon on the surface of a substrate.

As a method for producing micro-patterned surfaces[8], Nam and colleagues' approach could also benefit the design and production of materials that have new properties, and so find application in the fields of optics and microfluidics. The size of the channels made using the technique may be especially useful for microfluidic devices that manipulate single biomolecules. In some contexts, the authors' approach might even be suitable for producing simple, well-defined cavities. And because it is cheaper and simpler than other methods for fabricating channels, the technique is highly attractive for industrial applications.

The impact of this work[2] is not restricted to technology. Our understanding of crack dynamics will be enhanced through carefully designed experiments that use Nam and colleagues' tools. For instance, the dynamic evolution of competing cracks, and the 'refractory' transformation described by the authors, are intriguing problems that deserve to be studied. It will also be interesting to learn how the authors' techniques might be extended from the microscopic regime to control cracks at other spatial scales.

As with any technology, the method has some limitations — for example, the widths of the fabricated channels are highly dependent on the materials used to produce them. It may therefore be difficult to make cracks that have a wide range of depths or widths in a particular material, or to apply the technique to materials other than those presented in the current paper. Future investigations will surely offer fresh insight into how to overcome these difficulties. In the meantime, a diverse range of scientists will enjoy using this technique and will harvest the fruits that it provides. ■

Antonio J. Pons *is in the Department of Physics and Nuclear Engineering, Polytechnic University of Catalonia, Terrassa, Barcelona 08222, Spain.*
e-mail: a.pons@upc.edu

1. Lüth, H. *Solid Surfaces, Interfaces and Thin Films* (Springer, 2001).
2. Nam, K. H., Park, I. H. & Ko, S. H. *Nature* **485,** 221–224 (2012).
3. Ito, T. & Okazaki, S. *Nature* **406,** 1027–1031 (2000).
4. Broers, A. N., Hoole, A. C. F. & Ryan, J. M. *Microelectronic Eng.* **32,** 131–142 (1996).
5. Watt, F. et al. *Int. J. Nanosci.* **4,** 269–286 (2005).
6. Hakim, V. & Karma, A. *Phys. Rev. Lett.* **95,** 235501 (2005).
7. Chou, S. Y., Keimel, C. & Gu, J. *Nature* **417,** 835–837 (2002).
8. Kumar, G., Tang, H. X. & Schroers, J. *Nature* **457,** 868–872 (2009).

CARDIOVASCULAR BIOLOGY

Escaped DNA inflames the heart

High blood pressure can damage heart muscle cells and their mitochondrial organelles. DNA from degraded mitochondria has been shown to trigger inflammation leading to heart failure. SEE LETTER P.251

KLITOS KONSTANTINIDIS & RICHARD N. KITSIS

Heart failure is caused by weakening of the organ's muscle that results in ineffective pumping of blood around the body[1]. This disorder is a common result of diverse maladies, including heart attacks, high blood pressure and heart-valve problems. All these conditions generate mechanical and chemical signals that activate stress pathways, which ultimately cause heart-muscle dysfunction, cell death and alterations in the cells' milieu. This, in turn, results in changes in heart size and shape that lead over time to severe heart-muscle dysfunction and patient death. Cytokine proteins, produced by the cells of the immune system and by heart cells, play a part in these processes by promoting inflammation[2]. On page 251 of this issue, Oka *et al.*[3] describe a pathway that links the stresses that initiate heart failure with the production of cytokines in the muscle cells.

Mitochondria, the cell's energy-producing organelles, can suffer damage during the processes that ultimately lead to heart failure. Damaged mitochondria, or parts thereof, are often transported to cytoplasmic organelles known as lysosomes to be degraded by autophagy — a process by which cells break down and recycle their own components[4,5].

Because mitochondria originally evolved from bacteria, some of their molecular components are more similar to those of their bacterial relatives than to those of the cell's nucleus or cytoplasm. Specifically, nuclear DNA is modified by the addition of methyl groups on certain sequences known as CpG motifs, whereas most mitochondrial and bacterial DNA is not methylated. This differential feature allows cells of our immune system to recognize the DNA of invading bacteria: once lysosomes in the cells have engulfed the microbes, the receptor protein TLR9 senses the unmethylated CpG motifs and triggers the synthesis of proinflammatory cytokines[6,7].

Oka *et al.*[3] hypothesized that the DNA released during autophagy of mitochondria in heart muscle cells could trigger an inflammatory response similar to that generated by bacterial DNA during an infection. Furthermore, the authors speculated that a particular enzyme found in lysosomes, DNase II, could protect the heart from inflammation by digesting the mitochondrial DNA and therefore avoiding the inflammatory response. To explore these hypotheses, the researchers used a mouse model of heart failure in which the aorta (the main vessel that carries blood from the heart) was partially constricted. In such a model, the increased resistance against which the heart must pump (pressure overload) is akin to the stresses resulting from high blood pressure in humans.

The authors found that when mice deficient in DNase II were subjected to pressure overload, they had higher rates of heart failure and death than did control mice. These effects were ameliorated by pharmacological inhibition of TLR9 or by deletion of the TLR9-encoding gene. Oka *et al.* also induced high mortality and heart failure rates in wild-type mice by challenging them with a pressure overload of greater severity. Of note, inhibition or depletion of TLR9 in these mice also improved the deleterious effects of pressure overload, indicating that some mitochondrial DNA can escape lysosomal degradation even when DNase II is intact. Overall, the authors' results suggest that the sensing of mitochondrial DNA by TLR9 plays a part in promoting heart failure, whereas DNase II seems to have a protective role (Fig. 1).

Researchers have previously[8] shown that mitochondrial DNA released from dying cells into the extracellular space induces a TLR9-dependent inflammatory response in some cells of the immune system. What is new in Oka and colleagues' paper is the notion that mitochondrial DNA that escapes degradation can trigger a similar response in that same cell and that when this occurs in heart muscle cells, heart failure is exacerbated.

The current study engenders some interesting topics for future research. First, although the authors demonstrate the importance of the production of cytokines in heart muscle cells, non-muscle cells in the heart may also contribute to cytokine production if they are activated by mitochondrial DNA released from dying muscle cells. The relative contributions of these cytokine sources could be identified using tissue-specific deletion of the TLR9-encoding gene in mice. Second, given that autophagy of mitochondria is not directly demonstrated in this study, how does mitochondrial DNA end up in lysosomes? And, if autophagy of mitochondria is occurring, do other mitochondrial components have a role in this pathway?

A minority of patients with heart failure might have genetic mutations that lead to a loss of function of DNase II. However, whether the mechanism triggered by mitochondrial DNA is linked to most cases of heart failure is unclear. Indeed, it is surprising that ablation of TLR9 signalling results in such marked disease improvement in mice, given the many pathways that are activated during heart failure.

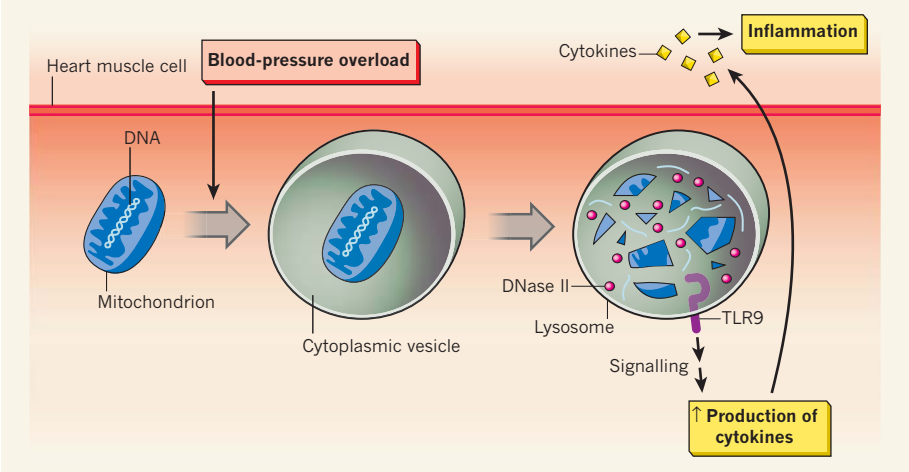


Figure 1 | A stressed heart cell. Oka *et al.*[3] carried out experiments in mice that suggest a possible mechanism contributing to heart failure. Various processes that put stress on the heart, such as blood-pressure overload, may damage mitochondria in heart muscle cells. These mitochondria are engulfed into cytoplasmic vesicles and transported to lysosomes — intracellular organelles in which mitochondrial components are digested by various enzymes such as DNase II, which degrades DNA. If mitochondrial DNA accumulates, then the protein TLR9 is activated. Signals from TLR9 induce the production of cytokine proteins, which are then secreted from the cell and act on the same and other heart cells to induce inflammation and contribute over time to adverse organ remodelling.

The results obtained by Oka *et al.* using mice suggest that the TLR9-mediated mechanism may be crucial for heart failure, but human studies will be needed to assess this. The pathway may be of greater importance in the adverse organ remodelling that immediately follows a heart attack, because such remodelling is characterized by more marked inflammation[9] than that taking place during chronic heart failure.

Moreover, the TLR9 pathway may be relevant to other disorders that involve inflammation in the absence of infection, such as atherosclerosis[10] and diabetes[11]. Multiple layers of complexity exist in the cytokine signalling involved in these diseases, and simultaneous inhibition of each would be a daunting task. Accordingly, if the pathway described by Oka and colleagues is found to have a key role in the development of these disorders, one can foresee the possibility of therapies based on small molecules directed against TLR9. ■

Klitos Konstantinidis *and* **Richard N. Kitsis** *are in the Departments of Medicine and Cell Biology and at the Wilf Family Cardiovascular Research Institute, Albert Einstein College of Medicine, Bronx, New York 10461, USA.*
e-mail: richard.kitsis@einstein.yu.edu

1. Hill, J. A. & Olson, E. N. *N. Engl. J. Med.* **358,** 1370–1380 (2008).
2. Mann, D. L. *Circ. Res.* **91,** 988–998 (2002).
3. Oka, T. *et al. Nature* **485,** 251–255 (2012).
4. Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. *Nature* **451,** 1069–1075 (2008).
5. Youle, R. J. & Narendra, D. P. *Nature Rev. Mol. Cell Biol.* **12,** 9–14 (2011).
6. Hemmi, H. *et al. Nature* **408,** 740–745 (2000).
7. Barton, G. M. & Kagan, J. C. *Nature Rev. Immunol.* **9,** 535–542 (2009).
8. Zhang, Q. *et al. Nature* **464,** 104–107 (2010).
9. Frangogiannis, N. G. *Circ. Res.* **110,** 159–173 (2012).
10. Libby, P., Ridker, P. M. & Hansson, G. K. *Nature* **473,** 317–325 (2011).
11. Hotamisligil, G. S. *Nature* **444,** 860–867 (2006).

CLIMATE SCIENCE

A grip on ice-age ocean circulation

Climate simulations based on an ocean model may hold the key to understanding why existing climate models have failed to deliver a clear picture of ocean circulation during the last ice age.

JOCHEM MAROTZKE

Simulating the climate of the most recent ice age is an important test for models that are used to project the climate of the future. But ice-age simulations have presented us with a persistent puzzle: simulations of a climatically crucial component of ocean circulation, the Atlantic meridional overturning circulation (AMOC), are inconsistent between models, and are often at odds with observational data. Writing in *Geophysical Research Letters*, Oka *et al.*[1] describe a series of climate simulations based on an ocean model that points to the cause of these discrepancies, and that offers an explanation as to why the ice-age AMOC was apparently much more prone to abrupt transitions than the modern one.

The AMOC consists of currents of warm water that, when summed up, flow northward in the upper 1,000 metres of the Atlantic Ocean, sink to great depths in the north, and then return southward as cold deep water (Fig. 1). The sinking is caused primarily by heat loss to the atmosphere, and occurs only if the water's salt concentration is high enough. This combination of warm water flowing northward and cold water flowing southward causes a considerable net transfer of heat northward. On release to the atmosphere, this heat contributes to the relatively mild European climate — an effect that establishes the AMOC as one of the most crucial oceanic contributors to modern climate. With respect to past climates, the AMOC is thought to have had a pivotal role in Earth's transition out of the most recent ice age into the current warm period[2].

There is robust observational evidence that at the height of the most recent ice age, the AMOC was less vigorous than it is today, and that its southward component occurred at shallower depths[3]. But ice-age simulations based on atmosphere–ocean climate models have so far shown a bewildering variety of AMOC strengths and structures[4,5]. Some models correctly show the ice-age AMOC as weaker than its modern-day equivalent; others exhibit no change from current conditions, and still others show an AMOC stronger than that observed today.

Oka and colleagues[1] suggest that this variety might be caused by the existence of a thermal threshold: if cooling in the North Atlantic were just a little stronger than it is today, the greater heat loss would cause AMOC strengthening; however, if North Atlantic cooling were stronger still, widespread sea-ice formation would ensue, insulating the ocean against further heat loss and weakening both the deep sinking and the AMOC. If the ice-age ocean was near this thermal threshold, small differences in the data entered into climate-model simulations could easily result in the models ending up on opposite sides of the threshold. This, in turn, would cause large inter-model differences in AMOC strength.

Oka *et al.* established the presence of the threshold using a clever analysis. They simplified the problem by using only the ocean component of their climate model, to which they applied different values for the amount of heat exchange — such as heat loss from the North Atlantic — with the atmosphere. Rather than using values corresponding to either ice-age or modern atmospheric conditions and thus heat exchange, the authors considered a mixed atmospheric state: 20% ice age and 80% modern, or 40% ice age and 60% modern, and so on. When the mix comprised no more than 40% ice-age heat exchange, their model generated a slightly stronger AMOC than that seen today, but with 60% or more ice-age heat loss, the AMOC weakened drastically. The full picture was more complex because surface winds, which were quite different in the ice-age simulations, had a strong influence on exactly where the threshold lay, but this simplified explanation captures the essential point.

Oka and colleagues' suggestion of a thermal threshold is striking in its simplicity, and is fully consistent with oceanographic folklore

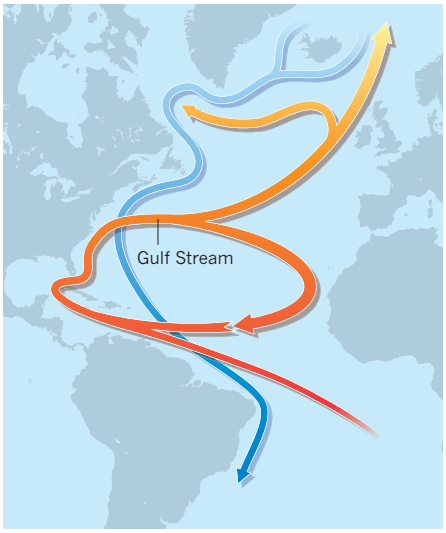


Figure 1 | The Atlantic meridional overturning circulation (AMOC). The AMOC is characterized by net northward flow of warm water (red) in the upper 1,000 metres of the Atlantic Ocean and southward flow of cold deep water (blue). The Gulf Stream is the most significant contributor to the northward warm-water flow. Oka *et al.*[1] describe a framework for understanding why the AMOC during the most recent ice age differs between existing climate models.

concerning the role of sea ice in suppressing deep-water sinking. The suggestion also offers an explanation as to why the AMOC was more likely to undergo abrupt transitions during the ice age than it seems to be today: if the ice-age AMOC was close to the thermal threshold, small changes in heat loss might have pushed it over this threshold. By contrast, today's conditions may lie far from the threshold, if the threshold even still exists.

Oka *et al.* have provided us with a conceptual framework for understanding why the ice-age AMOC differs between climate models. This understanding is a necessary first step towards making climate-model simulations consistent with observations. But the problem of the ice-age AMOC is not solved yet. First, we need confirmation that climate models with a weaker ice-age AMOC systematically show more sea-ice coverage in the northern deep-sinking regions than do those with a stronger ice-age AMOC. Second, the AMOC in ocean-only models such as that used by Oka and colleagues has long been known to be overly sensitive to minute details entered into the simulations[6]. The cause of this sensitivity is well understood[7] and suggests that simulations using atmosphere–ocean climate models are required to lend robustness to Oka and colleagues' interpretation. In these models, one would again apply the different percentage mixes of modern and ice-age inputs, but now the mix would be applied to greenhouse-gas concentrations, solar insolation and ice-sheet height. North Atlantic heat loss would then be calculated by the atmosphere–ocean model for every mix, rather than a heat-loss value being assigned to the ocean model as in Oka and colleagues' analysis. Because an atmosphere–ocean model is much harder to control than an ocean-only model, this approach is not guaranteed to produce the desired results. But it would be worth a try. ■

Jochem Marotzke *is at the Max Planck Institute for Meteorology, 20146 Hamburg, Germany.*
e-mail: jochem.marotzke@zmaw.de

1. Oka, A., Hasumi, H. & Abe-Ouchi, A. *Geophys. Res. Lett.* http://dx.doi.org/10.1029/2012GL051421 (2012).
2. Shakun, J. D. *et al. Nature* **484,** 49–54 (2012).
3. Lynch-Stieglitz, J. *et al. Science* **316,** 66–69 (2007).
4. Weber, S. L. *et al. Clim. Past* **3,** 51–64 (2007).
5. Otto-Bliesner, B. L. *et al. Geophys. Res. Lett.* **34,** L12706 (2007).
6. Weaver, A. J., Sarachik, E. S. & Marotzke, J. *Nature* **353,** 836–838 (1991).
7. Zhang, S., Lin, C. A. & Greatbatch, R. J. *J. Phys. Oceanogr.* **23,** 287–299 (1993).

STEM CELLS

One step closer to gut repair

The use of adult-tissue stem cells to treat gastrointestinal diseases holds much promise. A method for *in vitro* growth of gut stem cells and their use in repairing damaged intestines in mice has been described.

ANISA SHAKER & DEBORAH C. RUBIN

Adult-tissue stem cells have a property that makes them attractive to researchers in the emerging field of regenerative medicine. In contrast to most other cells from adult tissues, they can self-renew and generate the specific cell types that comprise the organs from which they are derived. Under the correct conditions, adult stem cells should be capable of generating fully functional, organ-specific tissues that would retain regenerative capacity and could restore the normal organ function of patients. In particular, stem-cell therapies might be beneficial for gastrointestinal disorders in which normal gut is lacking or damaged, as in short-bowel syndrome or inflammatory bowel disease. Now Yui and colleagues[1] describe in *Nature Medicine* how single stem cells from the mouse colon can be grown *in vitro* to produce mini-organs of epithelial tissue that can engraft onto a damaged intestine in mice.

The authors' work builds on the previous identification[2] of the Lgr5 protein as a marker for gut stem cells and on the establishment[3,4] of culture conditions that induce such cells to give rise to gut organoids — mini-organs that contain stem cells and other cell types that are typical of the intestinal epithelium. To test the possibility that gut stem cells could be used to regenerate intestinal tissue in live animals, Yui *et al.* isolated individual Lgr5-expressing stem cells from the colons of healthy mice (Fig. 1). These mice had been genetically engineered so that cells expressing Lgr5 also expressed a fluorescent protein, which facilitated their isolation. By culturing the individual stem cells, the researchers generated organoids and expanded them *in vitro*.

Yui *et al.* used the organoids to treat mice whose gut epithelial lining had been damaged

50 Years Ago

PROF. A. C. AITKEN, in a pamphlet entitled *The Case Against Decimalisation*, presents very skilfully the case against decimalization in general, starting with a lucid and cogent historical review ... He is less persuasive, however, in arguing for the duodecimal system, adoption of which in coinage, with a pound of twelve shillings, he suggests. He claims that this is a much more efficient system than the decimal system and that its adoption offers substantial advantages on practical as well as on arithmetical grounds.
From *Nature* 12 May 1962

100 Years Ago

One of the chief objections to the Daylight Saving Bill is the dislocation the scheme would effect in the zone system of time reckoning established by international conferences held successively in Rome and Washington thirty years ago. Mr. W. Ellis, F.R.S., refers particularly to this point in a short article in the March number of *The Horological Journal*. At present the prime meridian of Greenwich regulates the time of the civilized world. If the clocks of Great Britain are put forward one hour in summer, as proposed by the Bill, they will not show Greenwich time, but mid-European time; that is to say, our prime meridian, accepted by nations as regulating the time of the world, will be discarded by us for five months in every year ... An Act to enforce the alteration of clocks by putting them forward for one hour in the summer would introduce confusion in a scientific system and disturb accepted international standards. We cannot believe that such a proposal will ever be seriously entertained by Parliament.
From *Nature* 9 May 1912

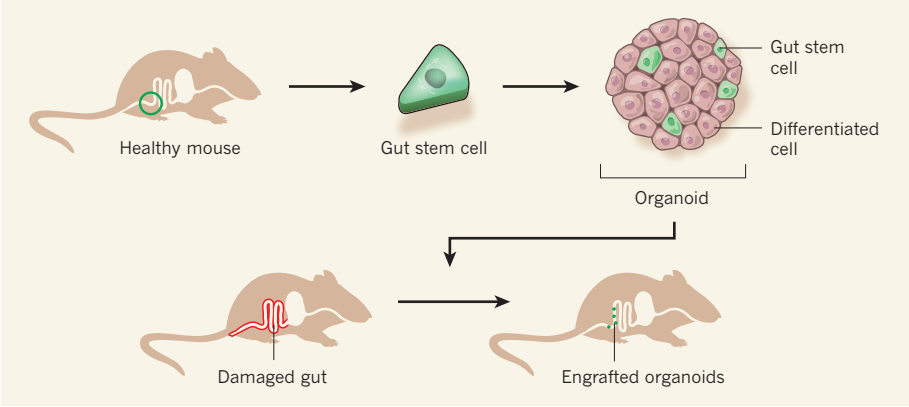


Figure 1 | Fixing a damaged intestine. Yui *et al.*[1] isolated individual stem cells from the gut of healthy mice and grew them *in vitro* using a medium that induced the cells to produce mini-organs (organoids) that contained stem cells and other types of differentiated cells commonly found in the colon. The authors expanded the organoids *in vitro* and then introduced them into mice whose intestines had been superficially damaged. Some of the organoids successfully attached to the damaged areas and survived for the six-month duration of the experiment.

in a way that resembles the effects of human ulcerative colitis — a form of inflammatory bowel disease. The authors introduced the organoids into the animals' intestine using an enema, which is a drug-delivery method often used in humans. Remarkably, the transplanted organoids were found to attach to the damaged areas and generated an epithelium that was reminiscent of that of normal colon and contained all the relevant cell types. Furthermore, because the donor Lgr5-expressing cells were fluorescent, the researchers were able to trace the engrafted tissue and determine that it was still self-renewing 25 weeks after treatment. The treated mice gained more weight than untreated animals during the experiment, suggesting possible disease amelioration.

The authors' findings[1] have exciting implications. It has been shown[5,6] that organoids isolated directly from normal intestine contain stem cells, can engraft onto damaged gut epithelium, and can support long-term epithelial regeneration. However, the current study goes one step further by showing that donor stem cells can be expanded *in vitro* to grow large numbers of organoids, which can then be transplanted and successfully engrafted into a host with a damaged gut. Therefore, using these techniques, stem cells from a patient's gut could be cultured in the laboratory to generate organoids that could then be used to provide functional intestinal tissue to the original donor. The stem cells could be isolated from samples taken during routine diagnostic procedures such as gastrointestinal endoscopy. What's more, the authors' study suggests that an enema-based system might be a viable mode of delivery.

However, much work remains to be done before a similar treatment in humans could be considered. In addition to weight gain, other parameters that are indicative of accelerated tissue healing should be monitored in future studies. Such parameters include the lack of diarrhoea, of blood in stools and of microscopic tissue damage. It will also be crucial to show that the transplanted epithelium not only is apparently intact but also functions normally. In Yui and colleagues' work, the organoids were engrafted only on areas of damaged tissue, which could be a limitation for the application of this approach in patients with short-bowel syndrome, a disorder in which areas of the small intestine are missing because of surgery or a birth defect. For these patients, engraftment onto the residual, normal intestine — in addition to generation of muscle, nerves and connective tissue — would be required.

Furthermore, the authors report a low rate of engraftment — this would need to be optimized to achieve reasonable therapeutic efficacy. Although it is encouraging that Yui *et al.* did not detect precancerous changes or polyps in the mice up to 25 weeks after transplantation, longer observation periods will be required to fully address the safety of the therapy. Despite these caveats, the study represents an important step towards realizing the promise of stem-cell-based therapies for gastrointestinal disorders. ■

Anisa Shaker *and* **Deborah C. Rubin** *are in the Department of Medicine, Division of Gastroenterology, Washington University School of Medicine, Saint Louis, Missouri 63110, USA.* **D.C.R.** *is also in the Department of Developmental Biology, Washington University School of Medicine.*
e-mail: drubin@dom.wustl.edu

1. Yui, S. *et al. Nature Med.* **18,** 618–623 (2012).
2. Barker, N. *et al. Nature* **449,** 1003–1007 (2007).
3. Sato, T. *et al. Nature* **459,** 262–265 (2009).
4. Sato, T. *et al. Gastroenterology* **141,** 1762–1772 (2011).
5. Booth, C., O'Shea, J. A. & Potten, C. S. *Exp. Cell. Res.* **249,** 359–366 (1999).
6. Tait, I. S., Evans, G. S., Flint, N. & Campbell, F. C. *Am. J. Surg.* **167,** 67–72 (1994).

ARTICLE

doi:10.1038/nature11019

ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner

Huai-Xiang Hao[1]*, Yang Xie[1]*, Yue Zhang[1]†, Olga Charlat[1], Emma Oster[1], Monika Avello[1], Hong Lei[1], Craig Mickanin[1], Dong Liu[1], Heinz Ruffner[2], Xiaohong Mao[1], Qicheng Ma[1], Raffaella Zamponi[1], Tewis Bouwmeester[2], Peter M. Finan[1], Marc W. Kirschner[3], Jeffery A. Porter[1], Fabrizio C. Serluca[1] & Feng Cong[1]

R-spondin proteins strongly potentiate Wnt signalling and function as stem-cell growth factors. Despite the biological and therapeutic significance, the molecular mechanism of R-spondin action remains unclear. Here we show that the cell-surface transmembrane E3 ubiquitin ligase zinc and ring finger 3 (ZNRF3) and its homologue ring finger 43 (RNF43) are negative feedback regulators of Wnt signalling. ZNRF3 is associated with the Wnt receptor complex, and inhibits Wnt signalling by promoting the turnover of frizzled and LRP6. Inhibition of ZNRF3 enhances Wnt/β-catenin signalling and disrupts Wnt/planar cell polarity signalling *in vivo*. Notably, R-spondin mimics ZNRF3 inhibition by increasing the membrane level of Wnt receptors. Mechanistically, R-spondin interacts with the extracellular domain of ZNRF3 and induces the association between ZNRF3 and LGR4, which results in membrane clearance of ZNRF3. These data suggest that R-spondin enhances Wnt signalling by inhibiting ZNRF3. Our study provides new mechanistic insights into the regulation of Wnt receptor turnover, and reveals ZNRF3 as a tractable target for therapeutic exploration.

Wnt proteins regulate the turnover of the transcription cofactor β-catenin and control key developmental gene expression programs[1,2]. Wnt proteins also induce planar cell polarity (PCP) or tissue polarity signalling, which governs cell and tissue movements[3]. Various secreted Wnt modulators and several feedback control mechanisms help to determine the proper signalling output. Perturbation of Wnt signalling can lead to degenerative diseases and cancer.

R-spondin proteins (RSPO1–4) strongly potentiate Wnt/β-catenin signalling and Wnt/PCP signalling[4–7] and regulate tissue patterning and differentiation[4,8–11]. R-spondin proteins are potent stem-cell growth factors[12]. RSPO1 strongly stimulates the proliferation of crypt stem cells[5,13] and protects mice from chemotherapy-induced mucositis[14]. Despite the biological and therapeutic significance, the mechanism of R-spondin action is unclear[4,7,15–17]. Recently, several groups including ours have demonstrated that the stem-cell marker LGR5 and its homologue LGR4 are R-spondin receptors essential for R-spondin-induced β-catenin and PCP signalling[18–20] (H.R. *et al.*, manuscript submitted). However, the mechanism by which R-spondin and LGR4 and LGR5 potentiate Wnt signalling remains unknown.

Here we identify the transmembrane E3 ubiquitin ligase ZNRF3 as the molecular target of R-spondin. Our data suggest that ZNRF3 inhibits Wnt signalling by promoting the turnover of frizzled and LRP6, and its activity is inhibited by R-spondin. Our study uncovers a new mechanism that controls Wnt receptor turnover, and its therapeutic exploration holds promise in regenerative medicine.

ZNRF3 as a negative regulator of Wnt pathway

Many negative regulators of Wnt/β-catenin signalling (for example, AXIN2) are β-catenin target genes and function in negative feedback loops. To identify new β-catenin target genes and potential negative regulators of Wnt signalling, primary tissue microarray data (NCBI accession GEO2109) were analysed for genes in which messenger RNA expression is positively correlated with *AXIN2* mRNA. This analysis identified *ZNRF3* and *RNF43* as β-catenin target genes. The expression of *ZNRF3* and *RNF43* was induced by Wnt3a conditioned media (Supplementary Fig. 1a). The expression of both genes was increased in primary colorectal tumours exhibiting hyperactive β-catenin signalling (Supplementary Fig. 1b). Furthermore, *ZNRF3* mRNA expression in the SW480 colorectal cancer cell line was down-regulated by β-catenin short interfering RNA (siRNA; Supplementary Fig. 1c).

ZNRF3 and RNF43 are highly related RING finger proteins (Supplementary Fig. 2). Both proteins contain a signal peptide, an extracellular domain, a transmembrane domain and an intracellular RING domain (Fig. 1a and Supplementary Fig. 2). ZNRF3 conjugated to green fluorescent protein (ZNRF3–GFP) is localized to the plasma membrane, whereas ZNRF3–GFP lacking the signal peptide (ZNRF3 ΔSP–GFP) is predominantly cytoplasmic (Supplementary Fig. 3a). These results were confirmed using a cell-surface protein biotinylation assay (Supplementary Fig. 3b). RNF43–GFP is also localized to the plasma membrane (data now shown). Furthermore, the intracellular fragment of ZNRF3 showed RING domain-dependent autoubiquitylation in an *in vitro* ubiquitylation assay (Supplementary Fig. 3c). These results suggest that ZNRF3 and RNF43 are E3 ubiquitin ligases localized to the plasma membrane.

We tested the function of ZNRF3 and RNF43 in Wnt/β-catenin signalling using HEK293 cells transfected with the Super-Topflash (STF) luciferase reporter. Depletion of ZNRF3, but not RNF43, strongly increased STF activity in the absence or presence of exogenous Wnt3a (Fig. 1b and Supplementary Fig. 4a). Overexpression of siRNA-resistant *ZNRF3* completely abolished *ZNRF3* siRNA-induced STF activation, suggesting that the effect of *ZNRF3* siRNA is on-target (Fig. 1c). In addition, the overexpression of wild-type ZNRF3 decreased Wnt3a-induced STF activation, whereas the overexpression of a ZNRF3 mutant lacking the RING domain (ZNRF3 ΔRING) strongly increased STF activity (Fig. 1c), indicating a

[1]Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA. [2]Novartis Institutes for Biomedical Research, Novartis Pharma AG, Postfach CH-4002 Basel, Switzerland. [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. †Present address: AstraZeneca, 35 Gatehouse Drive, Waltham, Massachusetts 02451, USA.
*These authors contributed equally to this work.

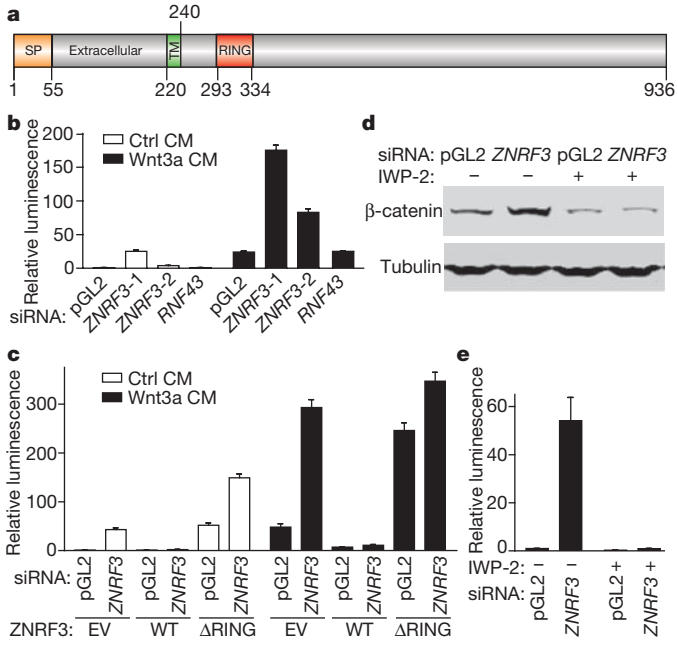


Figure 1 | ZNRF3 negatively modulates Wnt signalling. **a**, A schematic diagram of the domain structure of ZNRF3. SP, signal peptide; TM, transmembrane domain. **b**, Depletion of ZNRF3 increases the activity of the STF reporter in HEK293 cells. pGL2 siRNA acts as a negative control. CM, conditioned media; ctrl, control. **c**, *ZNRF3* siRNA-induced activation of STF is inhibited by siRNA-resistant *ZNRF3*, and ZNRF3 ΔRING increases STF activity. EV, empty vector; WT, wild type. **d**, *ZNRF3* siRNA-induced stabilization of cytoslic β-catenin is inhibited by the porcupine inhibitor IWP-2 (1 μM). **e**, STF reporter assay of HEK293 STF cells treated as in **d**. Error bars denote s.d.; $n = 4$ (**b**, **c**, **e**).

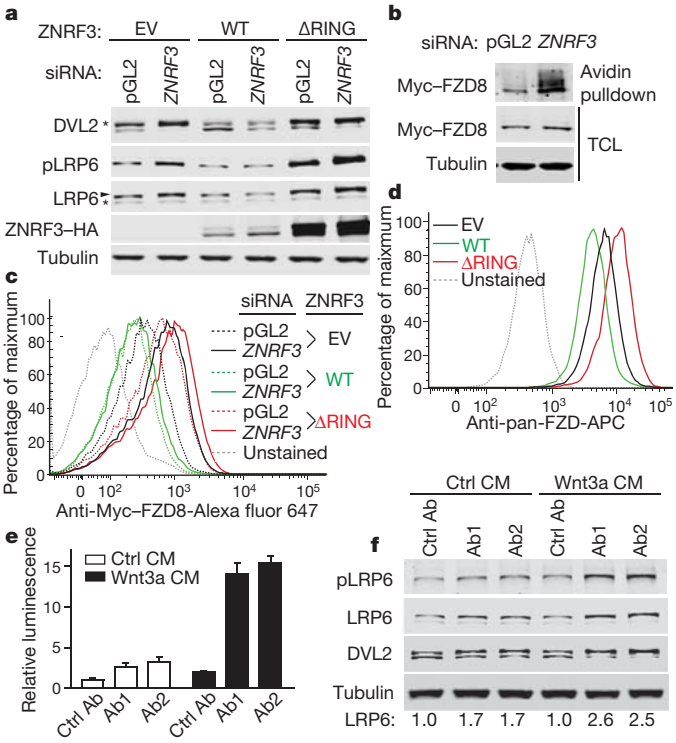


Figure 2 | ZNRF3 regulates the level of Wnt receptors on the cell surface. **a**, *ZNRF3* siRNA and ZNRF3 ΔRING increase the level of LRP6 and phosphorylated DVL2. The upper band of DVL2, indicated by an asterisk, is the phosphorylated form, and the lower band is the non-phosphorylated form. The mature and the ER form of LRP6 are indicated by an arrowhead and asterisk, respectively. pLRP6, phosphorylated LRP6. **b**, Depletion of ZNRF3 increases the cell-surface level of Myc–FZD8. TCL, total cell lysates. **c**, Flow cytometric analysis of membrane Myc–FZD8 in cells transfected with the indicated vector and siRNA. **d**, Overexpression of ZNRF3 decreases, and overexpression of ZNRF3 ΔRING increases, the cell-surface level of endogenous frizzled proteins. APC, allophycocyanin. **e**, ZNRF3 antagonistic antibodies (Ab1 and Ab2; 50 μg ml^{-1}) increase STF activity in HEK293 cells. Error bars denote s.d., $n = 4$. **f**, Immunoblots of indicated proteins for the same cells and treatments as described in **e**. Bottom, densitometric quantification of LRP6.

dominant-negative function of ZNRF3 ΔRING. Similarly, the overexpression of wild-type RNF43 blocked *ZNRF3*-siRNA-induced STF activation, whereas overexpression of RNF43 ΔRING increased STF activity (Supplementary Fig. 4b). These results indicate that ZNRF3 and RNF43 are functional homologues that act as negative regulators of Wnt/β-catenin signalling. On the basis of the threshold cycle (C_t) values observed in a quantitative PCR assay, *ZNRF3* is the dominantly expressed homologue in HEK293 cells (Supplementary Fig. 4a). Furthermore, IWP-2—a porcupine inhibitor that blocks Wnt secretion[21]—completely inhibited β-catenin accumulation and STF activation induced by *ZNRF3* siRNA or ZNRF3 ΔRING (Fig. 1d, e and data not shown) in the absence of exogenous Wnt3a. This result suggests that ZNRF3 suppresses the β-catenin signalling initiated by endogenous Wnt proteins, and this distinguishes ZNRF3 from other negative regulators of Wnt signalling such as adenomatous polyposis coli (APC), AXIN1/2 and glycogen synthase kinase 3-α/β (GSK3-α/β).

ZNRF3 regulates the stability of LRP6 and frizzled

Biochemical experiments were carried out to determine the molecular mechanism by which ZNRF3 regulates β-catenin signalling. Treatment with *ZNRF3* siRNA or overexpression of ZNRF3 ΔRING increased the levels of phosphorylated LRP6 and total LRP6 (Fig. 2a), and the effect of *ZNRF3* siRNA is blocked by the expression of siRNA-resistant *ZNRF3* (Fig. 2a). Increased LRP6 plasma membrane expression after ZNRF3 inhibition was confirmed using flow cytometry (Supplementary Fig. 5a).

Notably, treatment with *ZNRF3* siRNA or overexpression of ZNRF3 ΔRING increased the phosphorylation of dishevelled-2 (DVL2) (Fig. 2a), whereas DVL2 phosphorylation was decreased by ZNRF3 overexpression (Fig. 2a). As dishevelled phosphorylation is a direct readout of frizzled activation and is independent of LRP6 activation[22], these results suggest that the level or activity of frizzled might also be affected by ZNRF3. To test this hypothesis, HEK293 cells stably expressing amino-terminal Myc-tagged frizzled 8 (FZD8) were generated. In these cells, most Myc–FZD8 is trapped in the endoplasmic reticulum (ER), and only a small fraction is localized to the plasma membrane (data not shown). A cell-surface protein biotinylation assay showed that *ZNRF3* siRNA strongly increased the level of Myc–FZD8 on the plasma membrane without affecting the level of total Myc–FZD8 (Fig. 2b). Furthermore, *ZNRF3* siRNA and ZNRF3 ΔRING increased, whereas ZNRF3 decreased, the membrane level of Myc-FZD8 as determined by flow cytometry (Fig. 2c). Using pan-frizzled antibody 18R5 (ref. 23), the cell-surface level of endogenous frizzled proteins was found to be decreased or increased after the overexpression of ZNRF3 or ZNRF3 ΔRING, respectively (Fig. 2d). Inhibition of ZNRF3 also resulted in increased membrane levels of Myc–FZD4 (Fig. 3c), Myc–FZD5 (Supplementary Fig. 5b) and endogenous FZD6 (Supplementary Fig. 5c). Taken together, these results suggest that ZNRF3 regulates the membrane levels of frizzled and LRP6.

Because ZNRF3 is localized to the plasma membrane, antagonism of ZNRF3 using an antibody-based approach was explored. Two antibodies targeting the extracellular domain of ZNRF3 were identified that phenocopied *ZNRF3* siRNA treatment. Both antibodies enhanced Wnt3a-induced STF activity (Fig. 2e) and modestly increased the level of LRP6 (Fig. 2f) or membrane Myc–FZD8 (Supplementary Fig. 5d). These results further strengthen the conclusion that ZNRF3 inhibits Wnt signalling by decreasing the membrane levels of frizzled and LRP6.

Next we tested whether ZNRF3 inhibition affects the stability of LRP6 and Myc–FZD8. Overexpression of ZNRF3 ΔRING extended the half-life of LRP6 or membrane Myc–FZD8 in a ^{35}S-labelling-based or a cell-surface protein biotinylation-based pulse-chase experiment (Fig. 3a, b). These results suggest that ZNRF3 promotes the degradation of LRP6 and frizzled. Multiubiquitylation of frizzled promotes its lysosomal targeting and degradation[24]. We tested whether ZNRF3 is responsible for frizzled ubiquitylation. Myc-tagged FZD4 was used for this study because of its higher membrane expression level. Consistent with a previous report[24], FZD4 K0—in which all intracellular Lys residues are mutated to Ala—was expressed at a much higher level than wild-type FZD4 at the plasma membrane (Fig. 3c). Depletion of ZNRF3 increased the membrane level of wild-type FZD4, but not FZD4 K0 (Fig. 3c). Importantly, the overexpression of ZNRF3 ΔRING or depletion of endogenous ZNRF3 suppressed the ubiquitylation of FZD4 (Fig. 3d and Supplementary Fig. 5e), whereas the overexpression of ZNRF3 increased ubiquitylation of FZD4 (Supplementary Fig. 5f). Together, these results suggest that ZNRF3 promotes the degradation of frizzled by increasing its ubiquitylation.

Regulation of frizzled and LRP6 by ZNRF3 raises the possibility that ZNRF3 exists in the same complex as frizzled and LRP6. Indeed, co-expressed haemagglutinin-tagged ZNRF3 (ZNRF3–HA) and Myc–FZD8 can be co-immunoprecipitated (Fig. 3e). Furthermore, stably expressed ZNRF3–HA can be co-immunoprecipitated with endogenous LRP6 or FZD6 but not with other membrane proteins such as epidermal growth factor receptor (EGFR) and insulin receptor-β (IR-β) (Fig. 3f and Supplementary Fig. 6). Together, these results suggest that ZNRF3 exists in the same complex with frizzled and LRP6 and promotes their turnover.

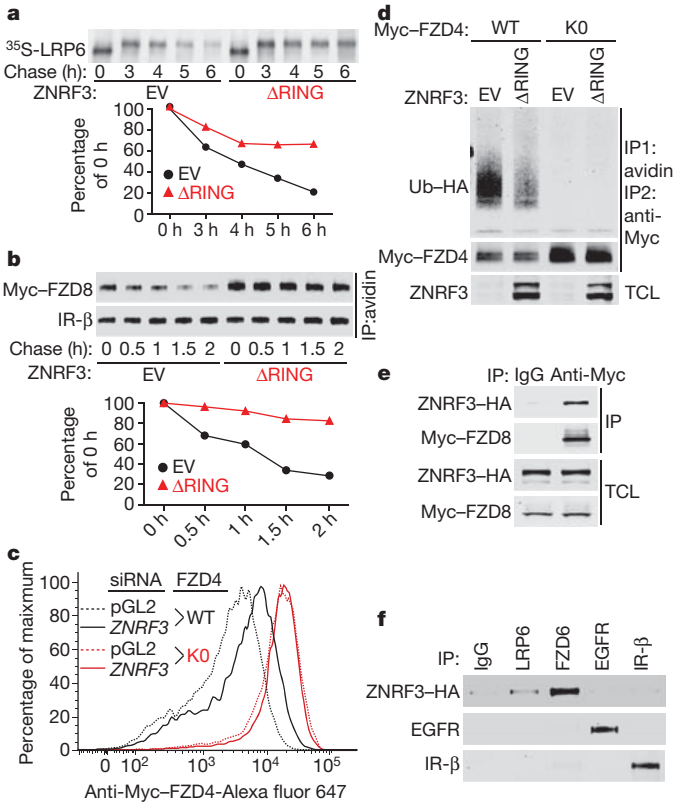


Figure 3 | ZNRF3 regulates the stability of LRP6 and frizzled through ubiquitylation. a, ZNRF3 ΔRING extends the half-life of LRP6 in a radioactive pulse chase assay. Bottom, densitometric quantification. **b**, ZNRF3 ΔRING extends the half-life of Myc–FZD8 in a surface protein biotinylation-based pulse chase assay. Insulin receptor-β (IR-β) acts as a control. Bottom, densitometric quantification of Myc–FZD8. IP, immunoprecipitate. **c**, ZNRF3 ΔRING increases the membrane level of Myc–FZD4 but not that of the K0 mutant. **d**, ZNRF3 ΔRING decreases ubiquitylation (Ub) of Myc–FZD4. **e**, Co-immunoprecipitation of Myc–FZD8 and ZNRF3–HA. **f**, Co-immunoprecipitation of ZNRF3–HA with endogenous LRP6 and FZD6. EGFR and IR-β act as negative controls.

R-spondin stabilizes Wnt receptors by suppressing ZNRF3

R-spondin potentiates Wnt/β-catenin and Wnt/PCP signalling through an unknown mechanism. Because frizzled is shared by Wnt/β-catenin and Wnt/PCP pathways and R-spondin induces dishevelled phosphorylation[16], we proposed that R-spondin might potentiate Wnt signalling by increasing the membrane level of frizzled proteins. Indeed, RSPO1 increased the membrane level of Myc–FZD8 (Fig. 4a and Supplementary Fig. 7a), and inhibited its degradation (Supplementary Fig. 7b). RSPO1 also increased the membrane level of endogenous frizzled proteins (Fig. 4b), Myc–FZD4 (Supplementary Fig. 7c), Myc–FZD5 (Supplementary Fig. 7d) and endogenous FZD6 (Supplementary Fig. 7e). RSPO2, RSPO3 and RSPO4 also increased the membrane level of endogenous frizzled proteins (Supplementary Figs 7f). Notably, RSPO1 decreased the ubiquitylation of FZD4 (Fig. 4c). In agreement with previous findings[17], RSPO1 and RSPO2 also increased the level of mature LRP6 (Fig. 4d). Consistent with a crucial role for LGR4 in R-spondin signalling, the depletion of LGR4 (Supplementary Fig. 8a) blocked RSPO1-induced accumulation of membrane Myc–FZD8 (Supplementary Fig. 8b, c) and LRP6 (Supplementary Fig. 8d), and abolished the effect of RSPO1 on FZD4 ubiquitylation (Supplementary Fig. 8e). Because ZNRF3 is required for frizzled ubiquitylation and R-spondin mimics *ZNRF3* siRNA in several assays, R-spondin might function by inhibiting ZNRF3. Indeed, RSPO1 did not increase STF activity in cells overexpressing ZNRF3 ΔRING (Supplementary Fig. 9a). In addition, RSPO1 did not further increase, and *LGR4* siRNA did not decrease, the membrane accumulation of Myc–FZD8 induced by ZNRF3 inhibition (Supplementary Fig. 9b, c). These results suggest that ZNRF3 functions downstream of R-spondin and LGR4.

Conservation of the extracellular domains (ECD) of ZNRF3 and RNF43 (Supplementary Fig. 2) and identification of ZNRF3 antagonistic antibodies indicate the existence of ligands for these proteins. Because R-spondin functionally mimics *ZNRF3* siRNA, the possibility that R-spondin directly interacts with ZNRF3 ECD was considered. Interestingly, RSPO1–GFP bound to cells overexpressing ZNRF3 ECD-TM, which lacks most of the intracellular domain, but did not bind to cells overexpressing FZD4 (Fig. 4e). This binding does not require endogenous LGR4 (Supplementary Fig. 10a). In a solution-based binding assay, RSPO1–GFP was copurified with ZNRF3 ECD–Fc, but not with FZD8 CRD (cysteine-rich domain)-Fc (Fig. 4f). As a control, Wnt3a was copurified with FZD8 CRD–Fc, but not with ZNRF3 ECD–Fc (Supplementary Fig. 10b). Furthermore, the mutation of ZNRF3 ECD Pro 103 to Ala severely disrupted its binding to RSPO1 in cell-based and solution-based binding assays (Fig. 4e, f). Together, these results indicate that R-spondin specifically interacts with the extracellular domain of ZNRF3.

If R-spondin increases Wnt signalling by binding to ZNRF3, the overexpression of ZNRF3 ECD might inhibit R-spondin-mediated signalling. We found that the overexpression of ZNRF3 ECD-TM completely blocked RSPO1-induced, but not Wnt3a-induced, STF activation and β-catenin stabilization (Fig. 4g, h), whereas ZNRF3 ECD-TM P103A had a much weaker effect (Supplementary Fig. 10c). Notably, the overexpression of ZNRF3 ECD-TM completely blocked RSPO1-induced membrane accumulation of frizzled proteins (Supplementary Fig. 10d). These results suggest that ZNRF3 ECD-TM inhibits R-spondin signalling, probably by functioning as a pseudoreceptor.

Because R-spondin binds to both LGR4 and ZNRF3, it might induce the association between the two proteins. Treatment with RSPO1 increased the interaction between exogenously expressed LGR4 and ZNRF3 in a coimmunoprecipitation assay (Fig. 5a). RSPO1 markedly

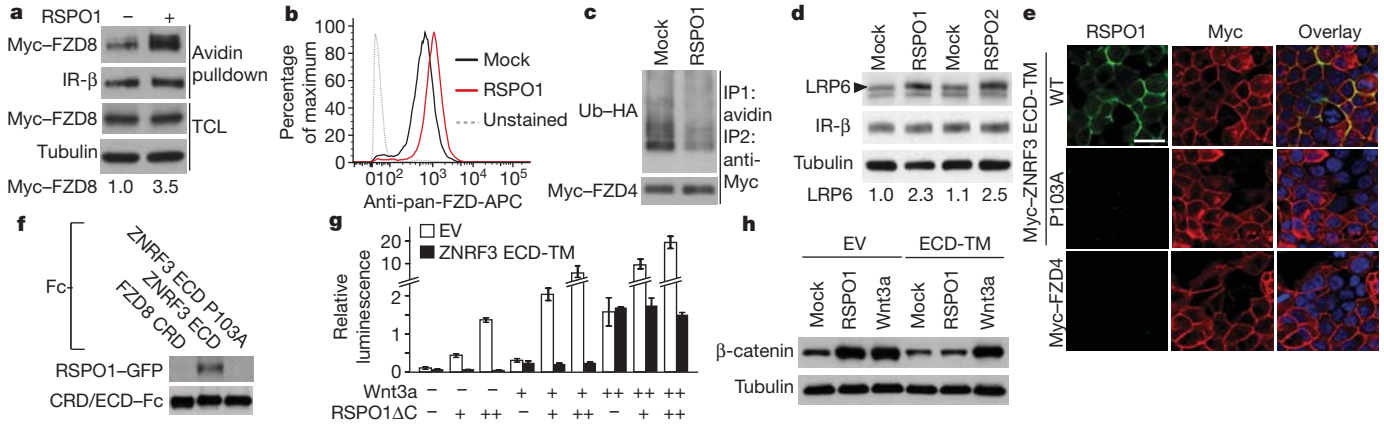


Figure 4 | RSPO1 increases the cell-surface level of frizzled proteins and functionally interacts with the extracellular domain of ZNRF3. **a**, RSPO1 (200 ng ml^{-1}) increases the membrane level of Myc–FZD8. Bottom, densitometric quantification. **b**, RSPO1 increases the membrane level of endogenous frizzled proteins. **c**, RSPO1 inhibits ubiquitylation of FZD4. **d**, RSPO1 and RSPO2 increase the level of LRP6. The mature form of LRP6 is indicated by an arrowhead. Bottom, densitometric quantification. **e**, RSPO1–GFP specifically interacts with overexpressed Myc-tagged ZNRF3 ECD-TM. The nuclei were stained with 4′,6-diamidino-2-phenylindole. Scale bar, 20 μm. **f**, RSPO1–GFP specifically interacts with the extracellular domain of ZNRF3. **g**, Overexpression of ZNRF3 ECD-TM specifically inhibits RSPO1-induced STF activation. RSPO1ΔC lacks 36 amino acid residues at the C terminus. Error bars denote s.d., $n = 4$. **h**, Overexpression of ZNRF3 ECD-TM specifically inhibits RSPO1-induced cytosolic β-catenin stabilization.

decreased the membrane level of ZNRF3 (Fig. 5b), but not that of ZNRF3 P103A (Supplementary Fig. 11a), in a cell-surface protein biotinylation assay. RSPO1-induced membrane clearance of ZNRF3 is LGR4-dependent (Supplementary Fig. 11b). As R-spondin induces the association between LGR4 and ZNRF3, we asked whether the forced dimerization of LGR4 and ZNRF3 would induce membrane clearance of ZNRF3. The heterodimerization domain DmrA or DmrC was fused to the carboxy terminus of ZNRF3 or LGR4, and these two chimaeric proteins were co-expressed in HEK293 cells. As seen in Fig. 5c, d, the treatment of cells with small molecule A/C dimerizer or with RSPO1 considerably reduced the membrane level of ZNRF3. Interestingly, RSPO1 and A/C dimerizer failed to decrease the membrane level of ZNRF3 ΔRING (Supplementary Fig. 11a, c), suggesting that the membrane clearance of ZNRF3 requires its E3 ligase activity. Taken together, these results suggest that R-spondin induces membrane clearance of ZNRF3 through LGR4, leading to accumulation of Wnt receptors on the cell surface.

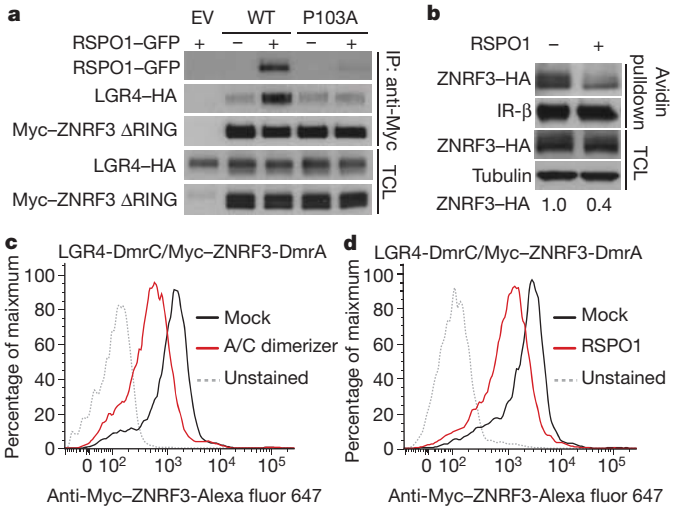


Figure 5 | RSPO1 increases the interaction between ZNRF3 and LGR4 and induces membrane clearance of ZNRF3. **a**, RSPO1 increases the interaction between ZNRF3 and LGR4. **b**, RSPO1 decreases the membrane level of ZNRF3. Bottom, densitometric quantification of cell-surface ZNRF3. **c**, **d**, Forced heterodimerization of LGR4 and ZNRF3 induces membrane clearance of ZNRF3.

ZNRF3 regulates β-catenin and PCP signalling *in vivo*

Because frizzled proteins are required for both Wnt/β-catenin and Wnt/PCP signalling, the inhibition of ZNRF3 is expected to promote both β-catenin and PCP signalling. We tested this hypothesis in various model organisms. The overexpression of human ZNRF3 ΔRING, but not wild-type ZNRF3, in zebrafish embryos resulted in the loss of anterior neural structures (Fig. 6a) and reduced expression of the anterior neural markers *hesx1* and *rx3* (Fig. 6b), consistent with ectopic activation of β-catenin signalling[25]. This phenotype was rescued by co-expression of AXIN1 (Supplementary Fig. 12a). Furthermore, the overexpression of human ZNRF3 ΔRING in *Xenopus* embryos led to axis duplication (Supplementary Fig. 12c) and increased expression of the β-catenin target gene *siamois* and *xnr3* in animal caps (Supplementary Fig. 12d). Induction of typical phenotypes associated with excessive β-catenin signalling by ZNRF3 ΔRING suggests that ZNRF3 suppresses β-catenin signalling *in vivo*. Precise regulation of PCP signalling output is required for normal gastrulation, and either increased or decreased PCP signalling disrupts convergent extension movements. The overexpression of wild-type ZNRF3 or ZNRF3 ΔRING in zebrafish embryos produced phenotypes characteristic of convergent extension defects, such as shortened body axis (Fig. 6c) and broader somites (Fig. 6d). Overexpression of wild-type ZNRF3 frequently caused axis bifurcation (Fig. 6d), and, interestingly, the same phenotype was also produced by overexpression of a dominant-negative frizzled[26]. Furthermore, the overexpression of wild-type ZNRF3 or ZNRF3 ΔRING disrupts dorsolateral cellular movements in a cell transplantation assay (Fig. 6e) and a fluorescent lineage tracing assay (Supplementary Fig. 12e). Notably, the overexpression of ZNRF3 ΔRINGΔSP (missing the signal peptide) had no effect on the expression of anterior neural markers and on convergent extension (Supplementary Fig. 12f)

To study the function of ZNRF3 in mice, *Znrf3* knockout mice were generated and backcrossed to a C57BL/6J background (Supplementary Fig. 13). *Znrf3*-deficient embryos died around birth. It is established that suppression of Wnt/β-catenin signalling in the lens placode is crucial for lens development; ectopic activation of β-catenin signalling in eyes blocks lens formation and eye-specific deletion of β-catenin leads to the formation of ectopic lentoid bodies[27–29]. The most obvious phenotype of *Znrf3* knockout embryos is the lack of lens formation (Fig. 6f and Supplementary Fig. 14a). Axin2–LacZ was negative in the lens placode of embryonic day (E)9.5 wild-type embryos, but it was positive in *Znrf3* knockout embryos (Fig. 6g). Expression of the β-catenin target gene

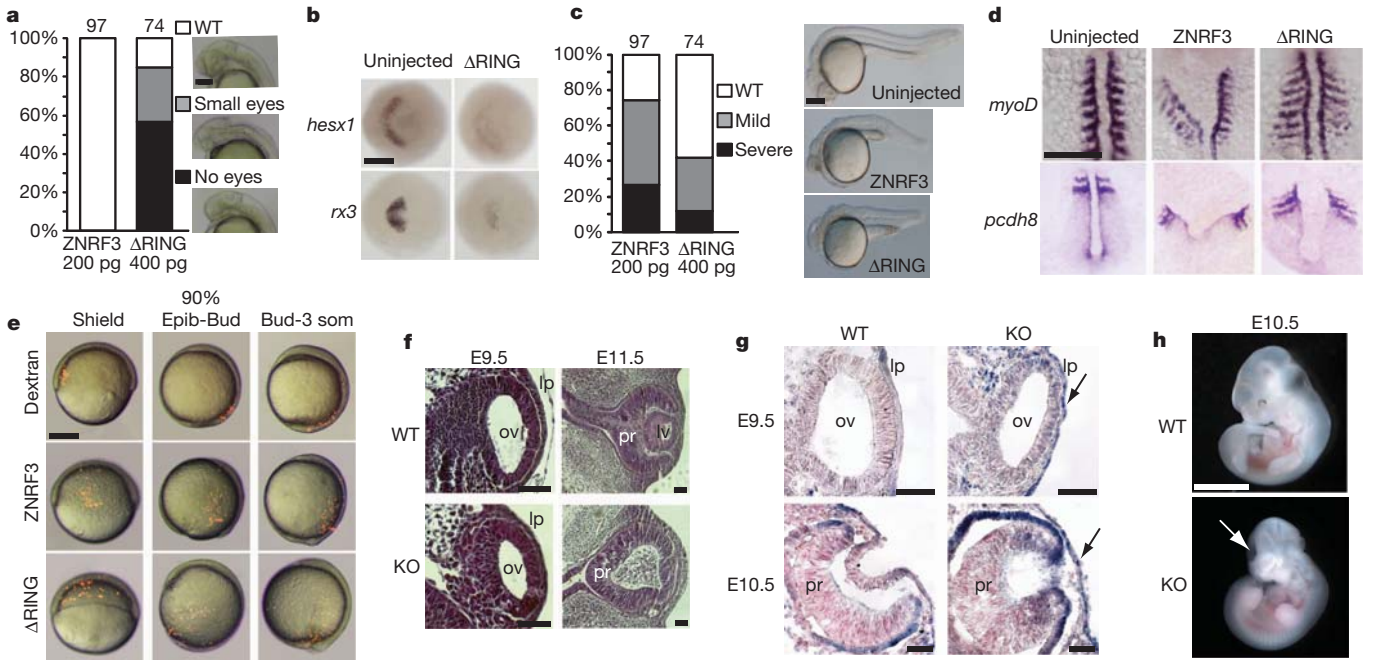


Figure 6 | ZNRF3 regulates both Wnt/β-catenin and Wnt/PCP signalling *in vivo*. **a**, Overexpression of ZNRF3 ΔRING in zebrafish embryos induces small eyes or loss of eyes, as shown in the right panel. Left, histogram of the frequency of the eye-loss phenotype. Numbers of embryos per treatment are indicated above the bars. **b**, Overexpression of ZNRF3 ΔRING in zebrafish greatly reduces the expression of anterior neural marker *hesx1* and the forebrain and retinal marker *rx3*. **c**, Overexpression of ZNRF3 or ZNRF3 ΔRING in zebrafish induces compressed somites and a shortened anterior–posterior axis. **d**, Top, *myoD* staining at the 8–9-somite stage. Zebrafish embryos injected with either mRNA have thinned and widened somites. Note that embryos injected with *ZNRF3* mRNA often display a bifurcated axis phenotype. Bottom, dorsal view of *pcdh8* expression at the 3-somite stage. **e**, Ovexpression of ZNRF3 or ZNRF3 ΔRING delays convergent extension movements. 90% Epib-Bud: 90% epiboly to tailbud stage; bud-3 som: tailbud to 3-somite stage. **f**, Haematoxylin and eosin staining of the eye region of E9.5 and E11.5 *Znrf3* wild-type (WT) and knockout (KO) mouse embryos. Lp, lens placode; lv, lens vesicle; ov, optic vesicle; pr, presumptive retina. **g**, Increased Axin2-LacZ activity in the lens placode of E9.5 and E10.5 *Znrf3* knockout mouse embryos. Arrows indicate Axin2-LacZ-positive cells. **h**, Gross morphology of E10.5 *Znrf3* wild-type and knockout mouse embryos with neural tube defects. Arrow indicates open neural tube in the *Znrf3* knockout embryo. Scale bars, 200 μm (**a**–**e**), 50 μm (**f**, **g**) and 2 mm (**h**).

Axin2 was also markedly increased in the eye region of E9.5 *Znrf3* knockout embryos (Supplementary Fig. 14b). In E10.5 *Znrf3* knockout embryos, thickening and invagination of presumptive retina is normal, but thickening and invagination of lens placode, which was associated with increased Axin2–LacZ, was completely blocked (Fig. 6g). These results are consistent with an important role of Wnt inhibition in lens development. Wnt/PCP signalling is essential for cell movements during narrowing of the folding neural plate[30]. Interestingly, about 20% of the *Znrf3* knockout embryos show neural tube closure defects (Fig. 6h and Supplementary Fig. 14a), which probably result from disrupted Wnt/PCP signalling. Taken together, these results suggest that ZNRF3 regulates both Wnt/β-catenin and Wnt/PCP signalling *in vivo*. Because RNF43 is a functional homologue of ZNRF3, the study of *Znrf3* and *Rnf43* double knockout mice should yield further information of these genes in embryonic development and adult tissue homeostasis.

Discussion

Our data support a model in which ZNRF3 and LGR4 form a receptor complex for R-spondin (Supplementary Fig. 15). In the absence of R-spondin, ZNRF3 ubiquitylates frizzled and promotes the degradation of frizzled and LRP6, leading to attenuated canonical and non-canonical Wnt signalling. When R-spondin is present, it induces the interaction between ZNRF3 and LGR4, leading to the membrane clearance of ZNRF3. This results in accumulation of frizzled and LRP6 on the plasma membrane and enhances canonical and non-canonical Wnt signalling. ZNRF3 is a unique transmembrane ubiquitin E3 ligase, the activity of which is directly regulated by ligand binding. We have demonstrated ZNRF3 as a tractable target through the identification of two ZNRF3 antagonizing antibodies. Potential application of such antibodies in regenerative medicine should be further explored.

Previously known cancer-associated Wnt pathway mutations all occur in the downstream components of the pathway, such as APC, AXIN1/2 and β-catenin. However, most Wnt inhibitors being developed at present would not inhibit Wnt signalling in tumours with downstream pathway mutations. During revision of the manuscript, RNF43 was identified as a tumour suppressor in cystic pancreatic tumours[31]. To our knowledge, RNF43 represents the first upstream Wnt pathway component mutated in cancers. This provides an exciting opportunity for the development of Wnt inhibitors. Various compounds such as porcupine inhibitor[21] and tankyrase inhibitor[32] should be tested for the treatment of cancers containing RNF43 mutations.

METHODS SUMMARY

Cell culture, transfection, STF assays, immunoprecipitation and immunoblotting were performed as previously described[32,33]. Information on plasmid expression constructs is available on request.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 October 2011; accepted 6 March 2012.
Published online 29 April 2012.

1. Clevers, H. Wnt/β-catenin signaling in development and disease. *Cell* **127,** 469–480 (2006).
2. MacDonald, B. T., Tamai, K. & He, X. Wnt/β-catenin signaling: components, mechanisms, and diseases. *Dev. Cell* **17,** 9–26 (2009).
3. Simons, M. & Mlodzik, M. Planar cell polarity signaling: from fly development to human disease. *Annu. Rev. Genet.* **42,** 517–540 (2008).
4. Kazanskaya, O. *et al.* R-Spondin2 is a secreted activator of Wnt/β-catenin signaling and is required for *Xenopus* myogenesis. *Dev. Cell* **7,** 525–534 (2004).
5. Kim, K. A. *et al.* Mitogenic influence of human R-spondin1 on the intestinal epithelium. *Science* **309,** 1256–1259 (2005).
6. Kim, K. A. *et al.* R-Spondin family members regulate the Wnt pathway by a common mechanism. *Mol. Biol. Cell* **19,** 2588–2596 (2008).

7. Ohkawara, B., Glinka, A. & Niehrs, C. Rspo3 binds syndecan 4 and induces Wnt/PCP signaling via clathrin-mediated endocytosis to promote morphogenesis. *Dev. Cell* **20,** 303–314 (2011).
8. Aoki, M. *et al.* R-spondin3 is required for mouse placental development. *Dev. Biol.* **301,** 218–226 (2007).
9. Blaydon, D. C. *et al.* The gene encoding R-spondin 4 (*RSPO4*), a secreted protein implicated in Wnt signaling, is mutated in inherited anonychia. *Nature Genet.* **38,** 1245–1247 (2006).
10. Kazanskaya, O. *et al.* The Wnt signaling regulator R-spondin 3 promotes angioblast and vascular development. *Development* **135,** 3655–3664 (2008).
11. Parma, P. *et al.* R-spondin1 is essential in sex determination, skin differentiation and malignancy. *Nature Genet.* **38,** 1304–1309 (2006).
12. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature* **469,** 415–418 (2011).
13. Ootani, A. *et al.* Sustained *in vitro* intestinal epithelial culture within a Wnt-dependent stem cell niche. *Nature Med.* **15,** 701–706 (2009).
14. Zhao, J. *et al.* R-Spondin1 protects mice from chemotherapy or radiation-induced oral mucositis through the canonical Wnt/β-catenin pathway. *Proc. Natl Acad. Sci. USA* **106,** 2331–2336 (2009).
15. Nam, J. S., Turcotte, T. J., Smith, P. F., Choi, S. & Yoon, J. K. Mouse cristin/R-spondin family proteins are novel ligands for the Frizzled 8 and LRP6 receptors and activate β-catenin-dependent gene expression. *J. Biol. Chem.* **281,** 13247–13257 (2006).
16. Wei, Q. *et al.* R-spondin1 is a high affinity ligand for LRP6 and induces LRP6 phosphorylation and β-catenin signaling. *J. Biol. Chem.* **282,** 15903–15911 (2007).
17. Binnerts, M. E. *et al.* R-Spondin1 regulates Wnt signaling by inhibiting internalization of LRP6. *Proc. Natl Acad. Sci. USA* **104,** 14700–14705 (2007).
18. Carmon, K. S., Gong, X., Lin, Q., Thomas, A. & Liu, Q. R-spondins function as ligands of the orphan receptors LGR4 and LGR5 to regulate Wnt/β-catenin signaling. *Proc. Natl Acad. Sci. USA* **108,** 11452–11457 (2011).
19. de Lau, W. *et al.* Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling. *Nature* **476,** 293–297 (2011).
20. Glinka, A. *et al.* LGR4 and LGR5 are R-spondin receptors mediating Wnt/b-catenin and Wnt/PCP signalling. *EMBO Rep.* **12,** 1055–1061 (2011).
21. Chen, B. *et al.* Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nature Chem. Biol.* **5,** 100–107 (2009).
22. Gonzalez-Sancho, J. M., Brennan, K. R., Castelo-Soccio, L. A. & Brown, A. M. Wnt proteins induce dishevelled phosphorylation via an LRP5/6- independent mechanism, irrespective of their ability to stabilize beta-catenin. *Mol. Cell. Biol.* **24,** 4757–4768 (2004).
23. Gurney, A. L. Frizzled-binding agents and uses thereof. US patent 201037041 (2011).
24. Mukai, A. *et al.* Balanced ubiquitylation and deubiquitylation of Frizzled regulate cellular responsiveness to Wg/Wnt. *EMBO J.* **29,** 2114–2125 (2010).
25. Kim, C. H. *et al.* Repressor activity of Headless/Tcf3 is essential for vertebrate head formation. *Nature* **407,** 913–916 (2000).
26. Nasevicius, A. *et al.* Evidence for a frizzled-mediated wnt pathway required for zebrafish dorsal mesoderm formation. *Development* **125,** 4283–4292 (1998).
27. Smith, A. N., Miller, L. A., Song, N., Taketo, M. M. & Lang, R. A. The duality of β-catenin function: a requirement in lens morphogenesis and signaling suppression of lens fate in periocular ectoderm. *Dev. Biol.* **285,** 477–489 (2005).
28. Kreslova, J. *et al.* Abnormal lens morphogenesis and ectopic lens formation in the absence of β-catenin function. *Genesis* **45,** 157–168 (2007).
29. Machon, O. *et al.* Lens morphogenesis is dependent on Pax6-mediated inhibition of the canonical Wnt/β-catenin signaling in the lens surface ectoderm. *Genesis* **48,** 86–95 (2010).
30. Wang, J. *et al.* Dishevelled genes mediate a conserved mammalian PCP pathway to regulate convergent extension during neurulation. *Development* **133,** 1767–1778 (2006).
31. Wu, J. *et al.* Whole-exome sequencing of neoplastic cysts of the pancreas reveals recurrent mutations in components of ubiquitin-dependent pathways. *Proc. Natl Acad. Sci. USA* **108,** 21188–21193 (2011).
32. Huang, S. M. *et al.* Tankyrase inhibition stabilizes axin and antagonizes Wnt signalling. *Nature* **461,** 614–620 (2009).
33. Zhang, Y. *et al.* RNF146 is a poly(ADP-ribose)-directed E3 ligase that regulates axin degradation and Wnt signalling. *Nature Cell Biol.* **13,** 623–629 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Yu, T. Lewis, Q. Song, J. Garver, J. Wang, B. Lu, B. Guo, Q. Fang, X. Shi, J. Sprunger and R.Freeman for technical assistance, R.-F. Kwong and T. Fleming for generating ZNRF3 antibodies, K. Lee and J. Halupowski for mouse maintenance, and J. Tchorz, A. Jaffe, N. Kubica, M. Hild, J. Solomon, Y. Yang, J. Tchorz, E. Wiellette, G. Michaud, D. Cutis and K. Seuwen for comments and advice. We also thank S. Goto for providing FZD4 and FZD4 K0 plasmids.

Author Contributions H.-X.H. initiated the project, characterized the function of ZNRF3 in cultured cells and mice, and identified ZNRF3 antagonistic antibodies. H.-X.H. and Y.X. discovered the R-spondin and ZNRF3 link. Y.X. led mechanistic studies on R-spondin, LGR4 and ZNRF3. H.-X.H., Y.X., Y.Z., H.L., C.M., D.L., H.R., X.M., Q.M., T.B., P.M.F., M.W.K., J.A.P., F.C.S. and F.C. conceived and designed the study. H.-X.H., Y.X., Y.Z., O.C., E.O., M.A., H.L., C.M., D.L., H.R., X.M., Q.M., R.Z., F.C.S. and F.C. designed and implemented experiments. H.-X.H., Y.X., Y.Z. and F.C. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to F.C. (feng.cong@novartis.com).

METHODS

Plasmids. Full-length human *ZNRF3* complementary DNA (NCBI accession number NM_001206998) was generated by fusing a short variant (NM_032173) and a synthesized 300-base-pair 5′ fragment. *ZNRF3* cDNA resistant to *ZNRF3*-1 siRNA was generated by two-step PCR and it was used as template for generating ZNRF3 ΔRING (missing amino acids 293–334) and ZNRF3 ECD-TM (missing amino acids 1–256). ZNRF3 was tagged with an N-terminal triple Myc epitope immediately after a signal peptide, or a C-terminal HA epitope. FZD8, FZD4 and FZD5 were tagged with an N-terminal triple Myc epitope right after the signal peptide. RSPO1 was fused with GFP at its C terminus. LGR4 was tagged with HA epitope at its C terminus. cDNAs are cloned in various mammalian expression vectors under control of the cytomegalovirus (CMV) promoter. Plasmids were sequenced to confirm the absence of undesirable mutations. Details of plasmids are available on request.

Cell culture, infection, transfection and RNA interference. HEK293 cells and its derivatives were grown in DMEM supplemented with 10% FBS. Various constructs were introduced into HEK293 or HEK293-STF cells through retroviral or lentiviral infection using standard protocols. Plasmid or siRNA transfection was done using FuGENE 6 (Roche) or Dharmafect 1 (Dharmacon), respectively. RSPO1, RSPO2, RSPO3 and RSPO4 were purchased from R&D Systems.

Sequences of siRNAs used are as follows: *ZNRF3-1* (QiagenSI03089744), sense, 5′-CCCAGUAUGAGACCAUGUATT-3′; antisense, 5′-UACAUGGUCUCAUACUGGGAG-3′. *ZNRF3-2* (Qiagen1027020), sense, 5′-GCUGCUACACUGAGGACUATT-3′; antisense, 5′-UAGUCCUCAGUGUAGCAGCCG-3′. *RNF43* (Dharmacon J-007004-09-0005), target sequence, 5′-GCAGAACAGAAAGCUAUUA-3′. *FZD6* (Dharmacon J-005505-07), target sequence, 5′-GAAGGAAGGAUUAGUCCAA-3′. *LGR4*-1 (Dharmacon J-003673-07), target sequence, 5′-AGGAUUCACUGUAACGUUA-3′ *LGR4*-2 (Dharmacon J-003673-08), target sequence, 5′-UUACUGAAGCGACGUGUUA-3′ *CTNNB1*, sense, 5′-UGUGGUCACCUGUGCAGCUdTdT-3′; antisense, 5′-AGCUGCACAGGUGACCACAdTdT-3′.

Luciferase assay. STF luciferase assays were performed using BrightGlo or DualGlo Luciferase Assay kits (Promega) according to the manufacturer's instructions.

Quantitative PCR with reverse transcription. Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen) and reverse transcribed with Taqman Reverse Transcription Reagents (Applied Biosystems) according to the manufacturer's instructions. The human colorectal cancer matched cDNA pair panel was purchased from Clontech. Transcript levels were assessed using the ABI PRISM 7900HT Sequence Detection System. Real-time PCR was performed in 12-μl reactions consisting of 0.6 μl of 20× Assay-on-Demand mix (premixed concentration of 18 μM for each primer and 5 μM for Taqman probe), 6 μl 2× Taqman Universal PCR Master Mix, and 5.4 μl diluted cDNA template. The thermocycling conditions used were 2 min at 50 °C, 10 min at 95 °C, followed by 40 cycles of 15 s at 95 °C and 1 min at 60 °C. All experiments were performed in quadruplicates. Gene expression analysis was performed using the comparative $\Delta\Delta C_T$ method with the housekeeping gene *GUSB* for normalization. The Assay-on-Demand reagents were purchased from Applied Biosystems.

Immunoblotting and immunoprecipitation. Immunoblotting and immunoprecipitation were performed as previously described[33]. Total cell lysates were prepared by lysing cells using RIPA buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS and 1 mM EDTA) supplemented with protease inhibitors and phosphatase inhibitors, followed by centrifugation at 20,000*g* for 10 min at 4 °C. Equal amount of proteins were resolved by SDS–PAGE, transferred to nitrocellulose membranes, and incubated with primary antibodies overnight at 4 °C. Secondary antibodies conjugated with either horseradish peroxidise or infrared dyes were used for signal visualization by ECL film or LI-COR Odyssey scanner, respectively. Quantification of immunoblotting bands was performed by densitometric analysis with AlphaEaseFC software. For co-immunoprecipitation experiments, cells were lysed in buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.8% Nonidet P40, phosphatase and protease inhibitors. Cleared cell lysates were incubated with the indicated antibodies and Protein G-sepharose beads (Amersham) overnight at 4 °C. Beads were washed four times with lysis buffer and the bound proteins were eluted in SDS sample buffer for immunoblotting analysis. The sources of primary antibodies are: anti-LRP6, anti-phospho-LRP6 (Ser 1490), anti-DVL2, anti-Myc tag, anti-GFP, anti-FZD6 (all Cell Signaling Technology); anti-IR-β, anti-Wnt3a (Abcam), anti-HA (Roche), anti-ZNRF3 (Santa Cruz), anti-β-catenin (BD Pharmingen) and anti-tubulin (Sigma).

Cell-based binding assay. HEK293 cells were transiently transfected with the indicated plasmids. Forty-eight hours after transfection, cells were incubated with RSPO1−GFP conditioned medium for 1 h. Cells were washed with PBS, fixed with 4% paraformaldehyde, incubated with anti-GFP and anti-Myc antibodies, stained with conjugated secondary antibodies and analysed by confocal fluorescence microscopy.

Solution-based binding assay. Conditioned medium of indicated proteins was mixed and incubated with Protein G beads overnight. Precipitates were washed extensively using buffer containing 50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.8% Nonidet P40, and phosphatase and protease inhibitors. Bound proteins were eluted in SDS sample buffer, and subjected to immunoblotting analysis.

Cell-surface protein isolation. Cell-surface proteins were isolated by whole-cell biotinylation and avidin agarose pull-down using Cell Surface Protein Isolation Kit (Pierce) according to manufacturer's instructions.

Flow cytometric analysis. Cells were collected using trypsin-free cell dissociation buffer (Invitrogen) and resuspended in FACS buffer (PBS with 1% BSA and 0.02% sodium azide). After blocking, cells were incubated with anti-LRP6 (R&D system), anti-Myc-Alexa fluor 647 (Cell Signaling Technology) or anti-pan-frizzled (18R5) antibody for 1 h at 4 °C, followed by incubation with conjugated secondary antibodies where applicable. After extensive washes using FACS buffer, cells were stained with propidium iodide and subjected to multichannel analysis using a BD LSR II flow cytometer. Fluorescence signals from propidium-iodide-negative cells were displayed in histogram plots. Inducible heterodimerization of LGR4 and ZNRF3 was achieved using an iDimerize Heterodimerization Kit (Clontech) by fusing DmrC and DmrA to the C terminus of LGR4 and ZNRF3, respectively.

Pulse-chase experiment. For ^{35}S-labelling-based pulse-chase, cells were starved for 1 h at 37 °C in labelling medium (methionine-free and cysteine-free DMEM supplemented with 10% dialysed FBS), and 200 μCi of ^{35}S-labelled methionine and cysteine (Perkin Elmer) was added to each 10-cm plate. After 1 h of labelling, radioactive labelling medium was replaced with chase medium containing 100 μg ml^{-1} methionine and 500 μg ml^{-1} cysteine. Cells were treated with 50% Wnt3a conditioned media 3 h after pulse labelling. At the indicated time points, cells were lysed in RIPA buffer and cell lysates were immunoprecipitated using an anti-LRP6 antibody at 4 °C overnight. After extensive washes using RIPA buffer, immunoprecipitates were treated with lambda phosphotase (New England Biolabs) to remove phosphates from LRP6. Bound proteins were eluted in SDS sample buffer and resolved by SDS–PAGE, and analysed by autoradiography. For biotinylation-based pulse-chase, cells were washed and incubated with PBS containing sulfo-NHS-SS-biotin (Pierce) for 15 min at 37 °C. After terminating the cross-linking reaction using quenching solution, cells were washed, fed with fresh growth medium and returned to the 37 °C incubator. At the indicated time points, cells were lysed in RIPA buffer, and biotinylated proteins were isolated using avidin agarose and subjected to immunoblotting analysis.

Cellular ubiquitylation assay. The cellular ubiquitylation assay was performed as described previously[24]. HEK293 cells were co-transfected with Myc–FZD4, HA–ubiquitin, and ZNRF3, ZNRF3ΔRING or empty vector. Thirty-six hours after transfection, cells were treated with biotinylation agent and lysed in RIPA buffer with 5 mM *N*-ethyl maleimide. Biotinylated cell-surface proteins were isolated by avidin agarose and eluted by boiling for 10 min in SDS lysis buffer (1% SDS, 50 mM NaF and 1 mM EDTA) supplemented with 50 mM dithiothreitol (DTT). Eluates were diluted using RIPA buffer, and immunoprecipitated using anti-Myc antibodies. Immunoprecipitates were resolved by SDS–PAGE and analysed by immunoblotting assay. To examine the effect of ZNRF3 knockdown on frizzled ubiquitylation, HEK293 cells were transfected with control pGL2 siRNA or *ZNRF3* siRNA and 48 h later transfected again with Myc–FZD4 and HA–ubiquitin. To examine the effect of R-spondin on frizzled ubiquitylation, HEK293 cells were transfected with Myc–FZD4 and HA–ubiquitin, and treated with 500 ng ml^{-1} RSPO1 overnight.

***In vitro* autoubiquitylation assay.** Recombinant N-terminal glutathione *S*-transferase-tagged ZNRF3 intracellular domain (ICD) protein was produced in *Escherichia coli* and purified using glutathione-agarose beads. Next, 0.6 μM of ZNRF3 ICD was incubated with 125 nM E1 ubiquitin-activating enzyme, 2 μM E2 ubiquitin-conjugating enzyme and HA-ubiquitin (Boston Biochem) at 37 °C for 6 h in ubiquitylation buffer (50 mM Tris-HCl, pH 8.0, 100 mM NaCl, 5 mM $MgCl_2$-ATP and 1 mM DTT). Samples were resolved by SDS–PAGE and subjected to immunoblotting analysis.

Generation of human ZNRF3 antibody. The HuCAL GOLD phage-display library was used for selection of ZNRF3-specific Fab fragments. Fc–ZNRF3 ECD (amino acids 56–216) protein produced in HEK293FS cells (Invitrogen) was used for phage panning. Fab clones were screened by ELISA using Fc–ZNRF3 ECD, and their binding to ZNRF3 was verified by FACS analysis using HEK293 cells stably expressing ZNRF3 ΔRING.

Zebrafish and *Xenopus* experiments. Zebrafish were maintained using standard methods[34,35]. Experiments using *Xenopus* embryos were performed as described previously[36]. *In vitro* transcription was performed to synthesize capped mRNA

using linearized plasmids containing human ZNRF3, ZNRF3 ΔRING and GFP as a template using a mMESSAGE mMACHINE kit (Ambion). For zebrafish, 200 pg of human *ZNRF3* wild-type mRNA or 400 pg of ZNRF3 ΔRING mRNA was injected into the embryos at the 1–2-cell stage. For *Xenopus*, 200 pg of *ZNRF3* wild-type, ZNRF3 ΔRING or control GFP mRNA was injected into two blastomeres in 4-cell-stage embryos at the marginal zone. For whole-mount *in situ* hybridization, embryos at the indicated stages were fixed overnight in 4% paraformaldehyde, and stained with digoxigenin (DIG)-labelled antisense probes using standard protocols[34]. To analyse the expression of *Xenopus znrf3*, total RNA was extracted from embryos at different stages, and analysed by quantitative PCR with reverse transcription (qRT–PCR) using Applied Biosystems SYBR-Green Master Mix. The primers used were: *znrf3*, 5′-GATGGAGAGGAG CTGAGAGTCATTC-3′ (forward), 5′-GATAACTCGCTGTTGCTGCTG-3′ (reverse); H4 histone, 5′- CGGGATAACATTCAGGGTA-3′ (forward), 5′-TCCATGGCGGTAACTGTC-3′ (reverse). Samples were normalized against H4 histone as an internal control. For RT–PCR with *Xenopus* animal caps, mRNA was injected into the animal poles of both blastomeres at the 2-cell stage. The animal caps were isolated at stage 8.5 and cultured until stage 10.5 for RT–PCR. The primers used were: *siamois*, 5′-CTCCAGCCACCAGTACCAGATC-3′ (forward), 5′-GGGGAGAGTGGAAAGTGGTTG-3′ (reverse); *xnr3*, 5′-TCC ACTTGTGCAGTTCCACAG-3′ (forward), 5-ATCTCTTCATGGTGCCTCAGG-3′ (reverse), and *EF-1α* (also known as *eef1a1*), 5-CAGATTGGTGCTGGATATGC-3′ (forward), 5′-ACTGCCTTGATGACTCCTAG-3′ (reverse). Cell transplantation experiments were performed essentially as described[34]. In brief, donor embryos were injected with the *ZNRF3* or ZNRF3(ΔRING) mRNA and a 4% tetramethylrhodamine-dextran (10 kDa) solution. Small clones of cells were transplanted from donor embryos at the high-to-oblong stages to stage-matched wild-type hosts, and these were then individually tracked and photographed using fluorescence microscopy. Analysis of convergent extension movements by cell tracking was performed as previously described[37]. In brief, 1 nl of 10 kDa dextran-conjugated Alexa 488 lineage tracer (Invitrogen) was injected into the yolk just below the margin at the 256-cell stage. The embryos were observed for cell movements towards the midline of the embryo and extensions along the anterior–posterior axis. Live images were taken at 30% epiboly, shield and 75% epiboly stages of the same embryos.

Generation and characterization of *Znrf3*-deficient mice. In the targeting vector, exon 7 encoding the RING domain is flanked by two *loxP* sites. Linearized targeting vector was electroporated into 129/SvJ embryonic stem (ES) cells, and G418-resistant ES clones were first screened by nested PCR, and then subjected to Southern blot analysis. Genomic DNA was digested with XmnI or BglII restriction enzymes, and hybridized with probes positioned outside the 5′ and 3′ homologous regions, respectively (Supplementary Fig. 13). ES clone 5A7 was used for blastocyst injection and chimaeric males were mated with Cre-deleter mice in the C57BL/6J background. F_1 mice with Cre-mediated deletion of exon 7 were identified by PCR, and further backcrossed in the C57BL/6J background. Wild-type, heterozygous and homozygous mice were identified by 'multiplex' PCR with the following three primers: forward primer 1, 5′-TATCATGGTCTGTATACCGGGATCG-3′; forward primer 2, 5′-CATACT TTGGGCTCATGAGCAAGC-3′; reverse primer, 5′- GCAGGTATACATTAC CACACCC-3′. Deletion of the RING domain of mouse Znrf3 creates a frameshift and premature termination. Truncated *Znrf3* transcript was not detected by qPCR assay in *Znrf3* knockout samples, so it is presumably degraded by non-sense-mediated mRNA decay. *Znrf3*$^{-/-}$ mouse embryos and wild-type littermate controls were generated by timed mating of heterozygous parents. At the indicated embryonic stage, pregnant females were euthanized and embryos were dissected out for imaging or histology analysis after fixation in 4% paraformaldehyde overnight at 4 °C. After dehydration in gradient serials of ethanol, the embryo was paraffin-embedded on head for horizontal sectioning, and slides were stained by haematoxylin and eosin. Whole mount *in situ* hybridization with E9.5 mouse embryos was carried out according to standard protocols using 25 nM double DIG-labelled locked nucleic acid probe from Exiqon. The mouse *Axin2* probe sequence was 5′-TCTCTAACATCCACTGCCAGA-3′. For the LacZ staining experiment, Znrf3 heterozygous males carrying the *Axin2*-LacZ transgene were mated with Znrf3 heterozygous females to generate E9.5 and E10.5 mouse embryos. Embryos were freeze-sectioned after fixation and cryopreservation in sucrose gradient. Frozen sections were stained with Tissue Stain Base Solution containing X-Gal (Millipore), post-fixed in 4% paraformaldehyde and counter-stained with 0.005% nuclear fast red. Primers for *Axin2*-LacZ genotyping were: 5′-AAGAAGAAGAGGAAGGTGGAAGATCCCGTCGTTTTAC-3′ (forward), 5′-GAGACGTCACGGAAAATGCCGCTCATC-3′ (reverse).

Statistical analysis. Results are expressed as mean ± s.d. from an appropriate number of experiments as indicated in the figure legends.

34. Nusslein-Volhard, C. & Dahm, R. *Zebrafish. A Practical Approach.* (Oxford Univ. Press, 2002).
35. Westerfield, M. *The Zebrafish Book: a Guide for the Laboratory Use of Zebrafish (*Brachydanio rerio*).* (Univ. Oregon Press, 1995).
36. Goentoro, L. & Kirschner, M. W. Evidence that fold-change, and not absolute level, of β-catenin dictates Wnt signaling. *Mol. Cell* **36,** 872–884 (2009).
37. Gerdes, J. M. *et al.* Disruption of the basal body compromises proteasomal function and perturbs intracellular Wnt response. *Nature Genet.* **39,** 1350–1360 (2007).

REVIEW

doi:10.1038/nature11117

Engineering the third wave of biocatalysis

U. T. Bornscheuer[1], G. W. Huisman[2], R. J. Kazlauskas[3,4], S. Lutz[5], J. C. Moore[6] & K. Robins[7]

Over the past ten years, scientific and technological advances have established biocatalysis as a practical and environmentally friendly alternative to traditional metallo- and organocatalysis in chemical synthesis, both in the laboratory and on an industrial scale. Key advances in DNA sequencing and gene synthesis are at the base of tremendous progress in tailoring biocatalysts by protein engineering and design, and the ability to reorganize enzymes into new biosynthetic pathways. To highlight these achievements, here we discuss applications of protein-engineered biocatalysts ranging from commodity chemicals to advanced pharmaceutical intermediates that use enzyme catalysis as a key step.

Biocatalysis is the application of enzymes and microbes in synthetic chemistry, and uses nature's catalysts for new purposes: applications for which enzymes have not evolved[1–5]. The field of biocatalysis has reached its present industrially proven level through several waves of technological research and innovations.

During the first wave of biocatalysis (Fig. 1), which started more than a century ago, scientists recognized that components of living cells could be applied to useful chemical transformations (in contrast to the fermentation processes, which had been commonplace for millennia already). For example, Rosenthaler synthesized (*R*)-mandelonitrile from benzaldehyde and hydrogen cyanide using a plant extract[6]; hydroxylation of steroids[7] occurring within microbial cells was also known. More recent examples are the use of proteases in laundry detergents[8], glucose isomerase to convert glucose to the sweeter-tasting fructose[9], and penicillin G acylase to make semisynthetic antibiotics[10]. The main challenge for these applications is the limited stability of the biocatalyst, and such shortcomings were primarily overcome by immobilization of the enzyme, which also facilitated the reuse of the enzyme.

During the second wave of biocatalysis, in the 1980s and 1990s, initial protein engineering technologies, typically structure based, extended the substrate range of enzymes to allow the synthesis of unusual synthetic intermediates. This change expanded biocatalysis to the manufacture of pharmaceutical intermediates and fine chemicals. Examples include the lipase-catalysed resolution of chiral precursors for synthesis of diltiazem (a blood pressure drug), hydroxynitrile-lyase-catalysed synthesis of intermediates for herbicides[11], carbonyl-reductase-catalysed synthesis of enantiopure alcohols for cholesterol-lowering statin drugs, lipase-catalysed synthesis of wax esters such as myristyl myristate or cetyl ricinoleate for cosmetics[12], and nitrile-hydratase-catalysed hydration of acrylonitrile to acrylamide for polymers[13] (where nitrile hydratase was obtained from whole cells of *Rhodococcus rhodochrous*). Apart from stabilization, the challenges now included optimizing the biocatalyst for the non-natural substrates.

The third, and present, wave of biocatalysis started with the work of Pim Stemmer and Frances Arnold in the mid and late 1990s. They pioneered molecular biology methods that rapidly and extensively

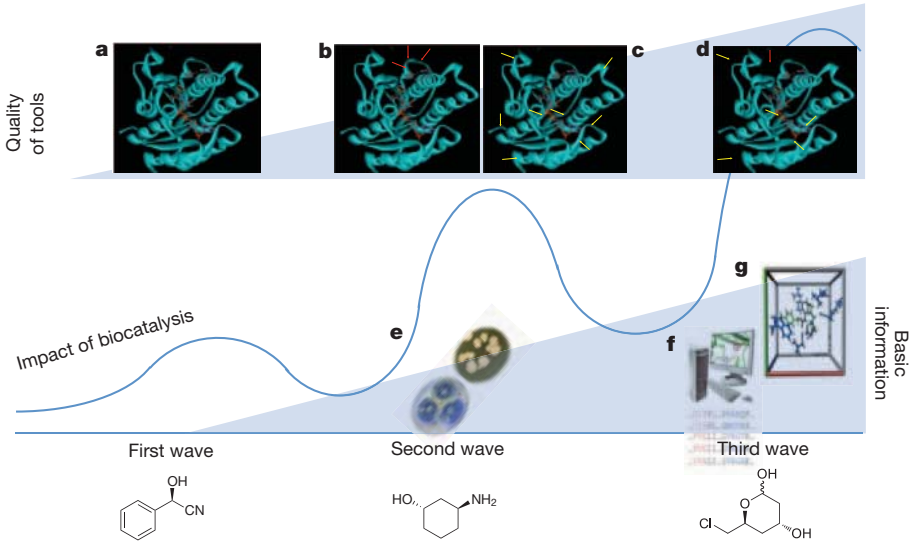


Figure 1 | The evolution of enzyme discovery and protein engineering strategies used to identify desired catalysts. Rational design (**b**) identifies distinct point mutations based on protein structures (**a**) or homology models, whereas random mutagenesis (**c**) combined with screening or selection is the basis for directed evolution experiments. Combining these methods makes it possible to create smaller, but smarter, libraries (**d**). The classical screening of enzymes by enrichment cultures (**e**) is now replaced by key motif database searches (**f**) to guide identification of novel enzymes or those with desired properties. Still in its infancy is the computational *ab initio* (or *de novo*) design of enzymes (**g**). The structures at the bottom refer to fine chemicals accessible through the different waves of biocatalysis. (*R*)-mandelonitrile (left) could already be obtained 100 yr ago from a plant extract; (1*S*,3*S*)-3-aminocyclohexanol (centre) is made by Novartis using an immobilized lipase; and 6-chloro-2,4,6-trideoxy-D-erythrohexapyranoside (right) is made by DSM in a process that requires an engineered aldolase to withstand high concentrations of acetaldehyde and to achieve high selectivity.

[1]Institute of Biochemistry, Department of Biotechnology & Enzyme Catalysis, Greifswald University, Felix-Hausdorff-Straße 4, D-17487 Greifswald, Germany. [2]Codexis Inc., 200 Penobscot Drive, Redwood City, California 94063, USA. [3]Department of Biochemistry, Molecular Biology and Biophysics, Biotechnology Institute, University of Minnesota, 1479 Gortner Avenue, Saint Paul, Minnesota 55108, USA. [4]Department of Chemical and Biological Engineering, Seoul National University, Seoul 151-744, Korea. [5]Department of Chemistry, Emory University, 1515 Dickey Drive, Atlanta, Georgia 30322, USA. [6]Merck Research Laboratories, Merck & Co., Rahway, New Jersey 07065, USA. [7]Lonza AG, Valais Works, CH-3930 Visp, Switzerland.

modify biocatalysts via an *in vitro* version of Darwinian evolution. The methods are now commonly called directed evolution, although this term has been in use since whole-cell experiments in 1972[14]. The initial versions of this technology involve iterative cycles of random amino-acid changes in a protein, followed by selection or screening of the resulting libraries for variants with improved enzyme stability, substrate specificity and enantioselectivity. Subsequent developments, discussed here, have focused on improving the efficiency of directed evolution to create 'smarter' libraries. Industrial-scale biocatalysis focused primarily on hydrolases, a few ketoreductases (KREDs), and cofactor regeneration and protein stability in organic solvents. In some cases, metabolic pathways were optimized; for example, combining genes from various natural strains to produce 1,3-propanediol (a monomer for polyesters) in a new host made it possible to switch from glycerol to the more convenient glucose as the feedstock[15].

As a result of the advances made during the present wave of biocatalysis, remarkable new capabilities can now be engineered into enzymes, such as the ability to accept previously inert substrates (a KRED for montelukast[16] or a transaminase for sitagliptin[17,18]) or to change the nature of the product that is formed (terpene cyclase variants that favour different terpenes[19] or amino-acid metabolism that makes alcohols for biofuels[20]). Novel enzymes are needed today to convert biomass into second- and third-generation biofuels[21,22], materials[23] and chemicals[24]. Key developments that enabled this third wave are advanced protein engineering[25–27] (including directed evolution), gene synthesis, sequence analysis, bioinformatics tools[28] and computer modelling, and the conceptual advance that improvements in enzymes can be more pronounced than previously expected. Engineered enzymes can remain stable at 60 °C in solutions containing organic solvents, can accept new substrates and can catalyse new non-natural reactions. This engineering may now take only a few months, thus greatly expanding the potential applications. In the past, an enzyme-based process was designed around the limitations of the enzyme; today, the enzyme is engineered to fit the process specifications.

About ten years ago, articles in *Nature*[29,30] and *Science*[31] reviewed the first and second waves of biocatalysis and provided a glimpse at what the third wave might bring. Today it is timely to assess the impact of this third wave and to speculate what the next decade might bring (Box 1). Although biocatalysis involves metabolic engineering[22,32,33] and synthetic biology, this Review focuses on enzymatic and whole-cell reactions.

Engineering enzymes to fit the manufacturing process

To minimize costs, chemical manufacture requires stable, selective and productive catalysts that operate under the desired process conditions. Engineering enzymes for such a process starts by defining the engineering goal, such as increased stability, selectivity, substrate range or, typically, a combination thereof. In 2000, before the third wave, only a few strategies were available to meet these goals. Enzyme immobilization could increase the stability of a protein, but the increases in stability were moderate and often insufficient for most chemical transformations. Directed evolution was also possible, but was still slow because it required construction and screening of large libraries that mostly contained variants with reduced, or even no, activity. Examples of drastic improvements were rarely of industrial relevance. The slow pace meant that the evolved proteins contained only a few changes and, thus, that the enzyme properties changed only slightly. Although several hundred enzymatic processes already had industrial uses[4], most involved enzymes and whole cells that had been marginally altered genetically[29].

In the past decade, our understanding of proteins and the number of available directed evolution strategies have both increased, making it possible to make large changes in enzyme properties. By and large, enzyme engineering continues to be a collection of case studies resulting from applying one of various possible approaches to the problem at hand, rather than there being a quantitative approach such as those used in disciplines such as civil, electric, software, or chemical engineering. Converting these case studies into engineering principles will require using free energy to connect the design goals to the structural changes needed (Fig. 2). Large changes in properties require large changes in free energy. For example, large changes in stability will require large free-energy changes in the folding–unfolding equilibrium. (Even irreversible protein unfolding starts with a reversible partial unfolding.) A molecular-level understanding of proteins suggests strategies that could be used for the improvements. For example, surface residues contribute to the folding–unfolding equilibrium and adding a proline residue in a loop lowers the entropy of the unfolded form. These strategies replace large libraries of random variants (mostly with poorer properties) with smaller, more focused protein libraries containing a high fraction of active and potentially improved variants (Fig. 1). Finally, by estimating the strength of various interactions (ion pairs on the surface or entropic contributions of adding a proline residue), researchers can estimate the changes needed to reach the goal. Few researchers explicitly use the free-energy-based measures to plan protein-engineering strategies today, but converting case studies into engineering principles requires a quantitative approach.

BOX 1

Requirements and examples of biocatalysis applications

- In traditional biocatalysis, natural products are converted into other natural products using natural reactions and pathways. Technical requirements: maintain microorganism cultures; conceptual requirements: possible to control natural biotransformations; examples: bread and cheese making, leather processing, beer and wine fermentation, and natural antibiotic production.
- In broad-substrate-range biocatalysis, chemical intermediates (non-natural products) are converted into other chemical intermediates using natural reactions and pathways. Technical requirements: use of defined enzymes (no interfering activity present); conceptual requirements: many enzymes have a broad substrate range; examples: manufacture of pharmaceutical intermediates using lipases and carbonyl reductases (alcohol dehydrogenases).
- In multistep biocatalysis, natural products are converted into fuels, materials and chemical feedstocks (non-natural products) using non-natural reactions and pathways. Technical requirements: protein engineering for major changes in stability, substrate range and type of reaction catalysed; conceptual requirements: enzymes can catalyse non-natural reactions and new combinations of enzymes create new pathways; examples: fuel molecules created using the isoprene biosynthesis pathways, amino-acid biosynthesis diverted to fuel alcohols.

New and improved methodologies

Over the past ten years, major advances in DNA technologies and in bioinformatics have provided critical support to the field of biocatalysis. These tools have promoted the discovery of novel enzymes in natural resources and have substantially accelerated the redesign of existing biocatalysts.

Advanced DNA technologies

Next-generation DNA sequencing technology has allowed parallel sequence analysis on a massive scale and at dramatically reduced cost. Whereas the cost of a human genome sequence analysis in 2002 was estimated at ~US$70,000,000, the price in 2012 has decreased more than 1,000-fold to less than US$10,000 (ref. 34), and LifeTechnologies, Illumina and Oxford Nanopore Technologies have announced that sequencing machines that are designed to sequence the entire human genome in a matter of hours will be available later in 2012 and will lower the cost per genome to less than US$1,000. Sequences of entire genomes from organisms from different environments, as well as environmental

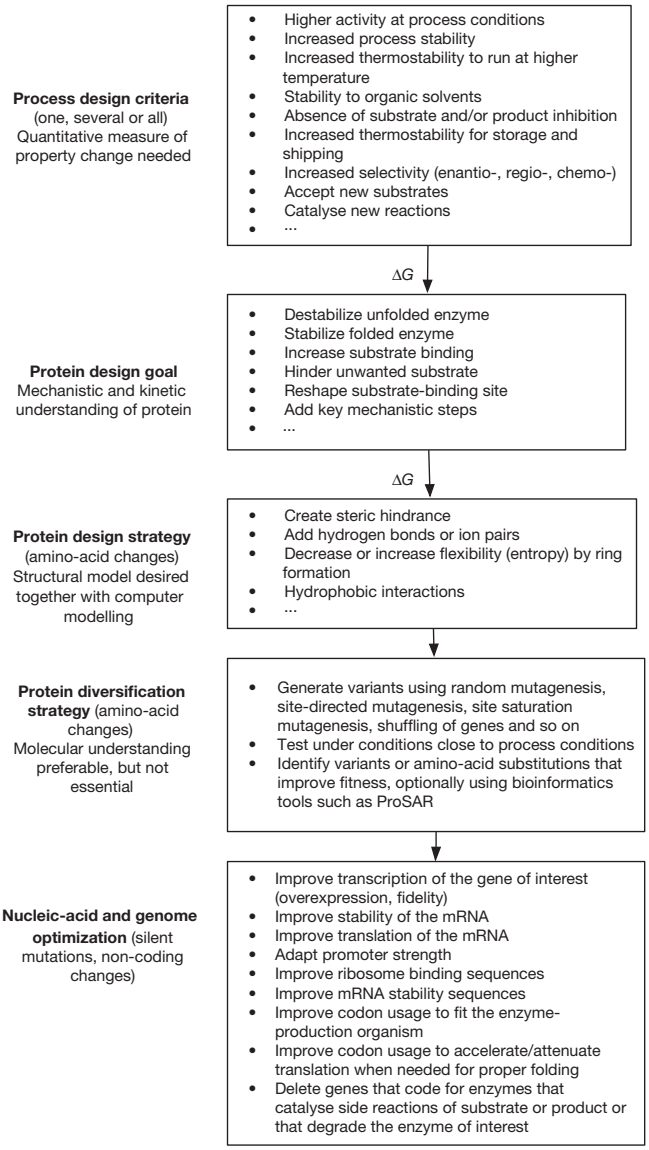


Figure 2 | Free energy (Δ*G*) connects design goals to the required structural changes via protein engineering strategies. This reasoning allows the use of more focused libraries. If the design goal lies far from the starting enzyme, then large changes in free energy are needed. Mechanistic and kinetic understanding of the protein unfolding and reaction mechanism identifies an engineering strategy to reach the design goal. Finally, analysis of the structure (which varies from qualitative inspection to extensive computer modelling) can identify the region or amino acids that must be changed. Goals that require large free-energy changes will likewise require more extensive changes in structure. mRNA, messenger RNA.

DNA samples that include unculturable organisms (metagenomes), have created a rich resource in which to search for novel biocatalysts[35], and will continue to do so. Massive high-throughput sequencing (>10,000,000 sequence reads) using the Illumina technology also facilitated the exploration and understanding of protein sequence–function relationships[36].

Low-cost DNA synthesis has replaced isolation of genomic DNA as the starting point for protein engineering. Whole-gene DNA synthesis further allows the codons to be optimized for the host organism and molecular architectural structures such as promoters, terminators, enhancers, restriction sites and so on to be introduced at convenient sites. This DNA synthesis uses traditional phosphoramidite chemistry, but optimized reaction conditions have improved coupling efficiency, increasing the overall quality and quantity of the polymer to make sequences even 200–250 nucleotides long. Parallel DNA synthesis using photolithographic and inkjet printing techniques further cut costs and speed synthesis[37]. DNA synthesis has been used to make entire sections of chromosomal DNA and even complete genomes for metabolic pathway engineering[38]. Whole-gene synthesis can also be used to make high-quality DNA libraries ranging from small, focused, site-saturation libraries to large, comprehensive gene collections. Customized genes and even gene libraries are becoming commodity chemicals similar to reagents and solvents found in today's research laboratories.

Novel tools in bioinformatics

Complementing the experimental advances, bioinformatics tools have become an integral part of modern protein engineering[39]. Multiple sequence alignments across large enzyme families and homology searches have identified genes with similar catalytic activities, leading to novel, potent biocatalysts[40]. The same sequence information allows the reconstruction of ancestral biocatalysts[40], which may have broader substrate range and catalytic promiscuity (see below). Multiple sequence alignments identify the most common amino acids at each position (the consensus sequence) and amino-acid substitutions that yield stable function enzymes. This data helps in the design of small libraries with a high proportion of catalytically active variants. These libraries have been used to discover biocatalysts with enhanced stabilities, catalytic functions and altered stereoselectivities[41].

Paralleling the advances in sequence-based protein engineering, structure-guided approaches have benefited from a rapid increase in protein structure coordinates deposited in the RCSB Protein Data Bank (http://www.pdb.org). Over the past decade, the repository has grown by over 450% to contain more than 77,000 protein structures. This facilitates both rational protein design and directed evolution, because structural alignment of related proteins helps to identify distinct similarities and differences guiding the more reliable design of mutant libraries.

The utility of smaller libraries was demonstrated in two different approaches to increasing the enantioselectivity of an esterase for resolution of methyl 3-bromo-2-methylpropionic acid, a chiral synthon[42]. Using random mutagenesis and screening 200 out of thousands of variants, the *E*-value for the enzyme (the selectivity of the enzyme for one enantiomer over the other) was increased from 12 to 19 (ref. 43). Recognizing that a relatively small increase (0.5 kcal mol^{-1}) in the difference in activation energy ($\Delta\Delta\Delta G^{\ddagger}$) for the two enantiomers was needed to generate a practical enzyme ($E > 30$), mutagenesis was focused at the active site. A library containing all possible single mutations at four positions (76 variants) yielded an enzyme with $E = 61$ ($\Delta\Delta\Delta G^{\ddagger} = 0.96$). Understanding the nature of the problem to be solved focuses the enzyme optimization approach on smaller libraries and gives bigger improvements. In this context, it is worth also mentioning a new method for continuous directed evolution using a combination of a phage infection system and a mutator plasmid in *Escherichia coli*[44].

Examples of engineered enzymes in industrial biocatalysis

As predicted by Schmid *et al.* in their forward-looking review in 2001[29], continuous regeneration of cofactors and a wider range of enzymes have been reported in the past ten years. However, the predicted applications of biochips and combinatorial biocatalysis have not yet materialized. The use of non-metabolizing cells for biocatalysis has proven to be more difficult than predicted and preference has instead shifted towards engineered enzymes used in crude and semipurified form. Whereas historically whole cells offered a simple and effective option for cofactor regeneration and enhanced enzyme stability, protein engineering and the use of single enzymes is now considered more economic and practical. The use of isolated enzymes have other advantages: they are easier to remove (less is added because they have more activity per unit mass), they tolerate harsher conditions, they eliminate potential diffusion limitations caused by cell membranes and they are easier to ship around the world. For example, KRED-based processes have now replaced whole-cell reductions and metal–ligand-based chemocatalysis, which were the industry standards during the past decade[45,46]. One exception is a whole-cell process to convert racemic hydantoins into

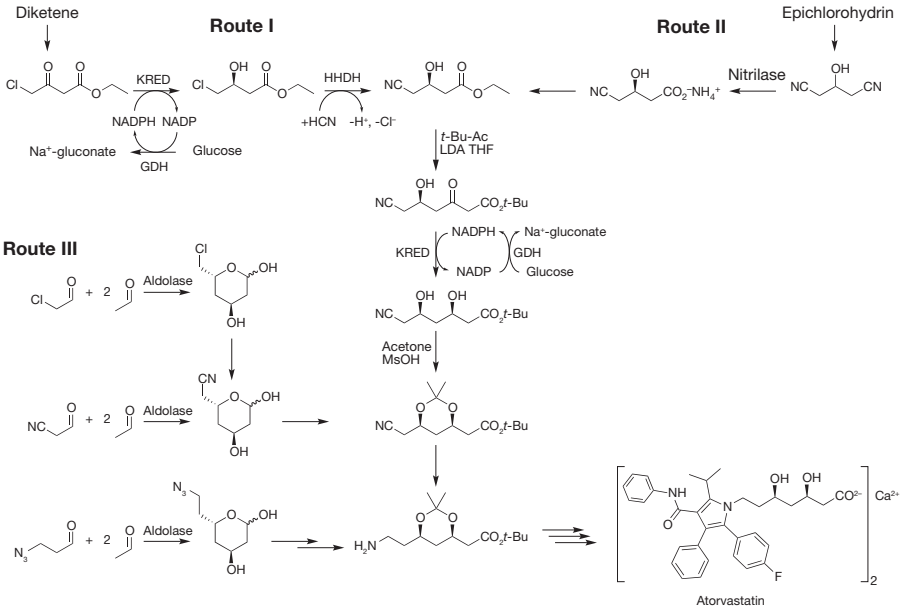


Figure 3 | Different enzymatic routes to the synthesis of the key side chain of atorvastatin (Lipitor). These processes use the combination of KRED with a halohydrin dehalogenase (HHDH) (route I), a nitrilase (route II) or an aldolase (route III). They differ not only in class of enzyme, but also in choice of (inexpensive) starting material, activity and selectivity of the biocatalyst, downstream processing, and yield and purity of final product. Routes I and II create one stereocentre, but route III creates both stereocentres required for the advanced pharmaceutical intermediate[52]. Introduction of the second chiral centre following routes I and II is also accomplished using KRED. LDA, lithium diisopropylamide; *t*-Bu, *tert*-butyl; THF, tetrahydrofuran.

optically pure, non-natural α-amino acids. Recombinant *E. coli* concurrently expressing hydantoinase, carbamoylase and racemase was found to be a simple and efficient production system, replacing the original process based on three immobilized enzymes in consecutive fixed-bed reactors[47,48]. Moreover, the whole-cell process required no metabolic flux controls and proceeded without undesired side reactions.

KREDs and other enzymes have been widely investigated for the manufacture of chiral intermediates for pharmaceuticals such as atorvastatin, the active ingredient in Lipitor, which is a cholesterol-lowering drug that had global sales of US$11,900,000,000 in 2010. Seven enzymatic approaches[2,49,50] (Fig. 3), differing not only in the choice of enzyme and starting material but also as to whether the product is a raw material (with a single chiral centre) or an advanced intermediate (with two chiral centres), have been developed. In all cases, success requires protein engineering to improve the reaction rate, the enantioselectivity, the stability to high substrate concentrations (up to 3 M, as in the nitrilase process[51]) or the stability to high solvent concentrations (20% butylacetate in the KRED process[52]). Apart from a highly active biocatalyst, a low-cost process also requires inexpensive raw materials and simple isolation of pure product in high yield. One current process leading to the advanced intermediate uses three biocatalytic steps: first, the combination of KRED and glucose dehydrogenase; second, the combination of this with a halohydrin dehalogenase to make the ethyl (*R*)-4-cyano-3-hydroxybutanoate intermediate (Fig. 3) at a rate of $>100\,t\,yr^{-1}$; and, third, the enzymatic reduction for the advanced diol intermediate[52].

Recent engineering[17] expanded the substrate range of transaminases to ketones with two bulky substituents. The enzyme engineering started with a small ketone substrate, created more space in the active site and used increasingly larger ketones. Several rounds of directed evolution increased the activity ~40,000-fold and yielded an engineered amine transaminase (Fig. 4) that can replace the transition-metal-based hydrogenation catalyst for sitagliptin manufacture. Starting from ATA-117, a close homologue of the wild-type enzyme, which had no detectable activity on the substrate, the first variant provided very low activity (0.2% conversion of $2\,g\,l^{-1}$ substrate using $10\,g\,l^{-1}$ enzyme) towards prositagliptin; the final variant converts $200\,g\,l^{-1}$ ketone to sitagliptin with 99.95% e.e. at 92% yield. The biocatalytic process not only reduced the total waste and eliminated all transition metals, but increased the overall yield and the productivity by 53% by comparison with the metal-catalysed process[18]. The numerous biocatalytic routes scaled up for pharmaceutical manufacturing (Table 1) demonstrate their competitiveness with traditional chemical processes.

Enzyme variants resulting from optimization studies are a unique source of starting points for future programmes, and by using the more stable enzymes engineered for one process the next optimization programme can be even faster. For instance, engineering KREDs to make R3HT (**3**) (see Table 1 for abbreviations and numbering of compounds) created many stable enzyme variants including some that were unsuitable due to low enantioselectivity. However, one of these unsuitable variants was the starting enzyme in engineering a KRED for DCFPE (**4**). One of the DCFPE enzymes was then the starting point for a montelukast (**5**) KRED, which in turn was a starting point for the duloxetine (**6**) KRED. Similarly, the transaminases generated during the evolution of the prositagliptin (**18**) transaminase can make other amines and may serve as starting points for new engineered enzymes for amine synthesis. Starting from a non-natural stabilized enzyme variant that already works in one process thus accelerates catalyst and process development in unprecedented ways.

Enzymatic conversions that simultaneously set two stereocentres are especially efficient ways to make complex molecules. For example,

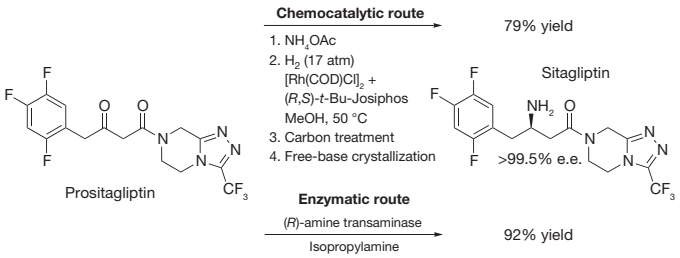


Figure 4 | Biocatalysis advances synthetic chemistry. The enzymatic route for the synthesis of sitagliptin using an engineered amine transaminase is superior to the chemical hydrogenation process, resulting in a higher yield of >99.5% e.e. optically pure product, higher productivity, reduced total waste and elimination of a transition-metal catalyst[17,18]. atm, atmospheres; e.e., enantiomeric excess; Me, methyl.

reduction of a ketone catalysed by a KRED sets the alcohol stereocentre. However, if a second stereocentre next to the ketone carbonyl racemizes rapidly in solution and the KRED is highly selective for one configuration, then the reduction reaction can set two stereocentres in one step. Examples include a penem intermediate (**8**), pseudoephedrine (**12**) and phenylephrine (**13**), as well as similar processes for chiral amines (**20**). Also, aldolases are now being used for subsequent reaction steps to generate multiple chiral centres, for example in the synthesis of statin intermediates (**33**). The single-enzyme cascade reactions as catalysed by aldolases have been further expanded to multi-enzyme cascade processes, for instance in the synthesis of 2′-deoxyinosine (**34**) *in vitro* or of complex molecules such as artemisinin (**42**) or Taxol *in vivo*.

Environmental advantages of biocatalytic processes

In the context of concerns about the environmental aspects of chemical manufacturing, biocatalysis provides an attractive alternative. The US Environmental Protection Agency awards five prizes each year in the Presidential Green Chemistry Award Challenge. The nominations emphasize the 12 Principles of Green Chemistry[53], which consider environmental factors as well as use of renewable feedstocks, energy efficiency and worker safety. Biocatalysis, using either enzyme technology or whole cells, has won 16 awards since 2000 (Table 2). Biocatalysts are made from renewable sources and are biodegradable and non-toxic, and their high selectivities simplify reaction work-ups and provide product in higher yields. Biocatalytic processes are also safe as they typically run at ambient temperature, atmospheric pressure and neutral pH. Hence, it is not surprising that so many of the awards go to biocatalysis.

The broad range of awards in Table 2 shows the application of biocatalysts beyond the pharmaceutical industry. Several applications involve polymers, especially polyesters. The optimized fermentation of lactic acid is the basis for a polylactic acid plant in Nebraska with a capacity of 141,000 tonnes per year, and a new non-natural metabolic route allows synthesis of 1,3-propanediol for the manufacture of SORONA polymer. The 1,4-butanediol fermentation yields a component of another polymer, spandex. In the synthesis of polyhydroxyalkanoates, the biocatalyst catalyses not just the monomer synthesis but also its polymerization. Yang and co-workers recently engineered these polyhydroxyalkanoate synthases to polymerize lactic acid into polylactic acid[54]. Several other awards involve new metabolic pathways to manufacture biofuels. For example, LS9, Inc. engineered *E. coli* bacteria to produce biodiesel. Adding the genes for plant thioesterases to *E. coli* diverted normal fatty-acid biosynthesis into synthesis of several fatty acids suitable for biodiesel. Then genes were added for enzymes to make ethanol and an enzyme to couple the ethanol and fatty acids to make fatty-acid ethyl esters, which can be used for biodiesel[22]. The amount of biodiesel produced is at least tenfold too low for the process to be commercially viable, but further engineering will probably increase the yield.

New concepts in protein engineering

Large changes in enzyme properties usually require multiple amino-acid substitutions because they make larger changes to the protein structure. However, simultaneous multiple amino-acid substitutions create exponentially more variants for testing. There are 7,183,900 possibilities for two substitutions anywhere in a 200 amino-acid protein and 9,008,610,600 possibilities for three substitutions. Many of these variants are inactive, and either all are created and tested to find the improved variants, or the library is screened only partially and incremental improvements in subsequent rounds of evolution are required.

The simplest solution to this problem is more efficient screening. Changes in substrate specificity may be monitored by high-throughput methods, such as fluorescence-activated cell sorting[55–57], which can screen tens of millions of variants in a short amount of time. Whittle and Shanklin made six simultaneous substitutions in the active site of a desaturase and then screened for growth on a different substrate. Only those variants with altered substrate specificities could grow[58]. Seelig and Szostak used very large random libraries (up to 10^{13} variants), from which they could select variants that catalysed an RNA ligation[59] based on binding of the product, but not the starting materials, to a column.

At present, the best approach to creating multiple mutations is to add them simultaneously but to limit the choices using statistical or bioinformatic methods. One statistical correlation approach is based on the ProSAR (protein structure activity relationship) algorithm used by Codexis researchers to improve the reaction rate of a halohydrin dehalogenase >4,000-fold[60]. Researchers made random amino-acid substitutions (an average of ten) in the dehalogenase and measured the rate of catalysis by the variants. Then statistical methods identified whether a particular substitution was beneficial. For example, variants that contained a Phe 186 Tyr substitution were, on average, better than those that did not. Some variants that contained such a substitution were not beneficial, owing to the detrimental effects of other mutations, but the statistical analysis identified that, on average, Phe 186 Tyr is a beneficial mutation. The final improved enzyme contained 35 amino-acid substitutions among its 254 amino acids.

γ-Humulene synthase catalyses the cyclization of farnesyl diphosphate via cationic intermediates to γ-humulene in 45% yield, but forms 51 other sesquiterpenes in smaller amounts. Keasling and co-workers substituted amino-acid residues in the active site stepwise and identified the contribution of each one to the product distribution[61]. Substitutions were combined to favour formation of one of the other sesquiterpenes. For example, one triple substitution created an enzyme that formed 78% sibirene; the natural enzyme forms 23% sibirene.

Another approach is to limit the location of changes to the active site and the type of changes to those known from sequence comparisons to occur often at these sites. Jochens and Bornscheuer used this approach to increase the enantioselectivity of a *Pseudomonas fluorescens* esterase. There were 160,000 (20^4) ways simultaneously to vary the four amino acids adjacent to the substrate in the active site. The researchers aligned the amino-acid sequences of >2,800 related enzymes to identify which amino acids are most common at these locations. This analysis limited the possibilities to several hundred variants, which were tested to find a double and a triple mutant with the desired selectivities[41]. Another important advance that allows multiple mutations is the recognition that mutations often destabilize proteins[62–64] and that starting with a very stable protein therefore allows it to tolerate a greater number and range of changes[65,66].

Because the workload for screening larger libraries containing multiple mutations increases exponentially with library size, most researchers work on the assumption that beneficial mutations are mostly additive[67] and that synergistic effects are rare, except for nearby changes. Indeed, combining beneficial single mutations often yields additive improvements (for example increasing the stability of an esterase from *Bacillus subtilis* to an organic solvent[68] or increasing the enantioselectivity of a lipase from *Pseudomonas aeruginosa*[69]). However, often the contributions do not add up exactly or have unexpected behaviour. For example, mutations A and B by themselves may be deleterious, but together they may be beneficial. Weinreich and co-workers[70] investigated such cooperative interactions in the evolution of a β-lactamase with higher activity. Mutation A increased the reaction rate but destabilized the β-lactamase. The overall effect was slightly beneficial. Mutation B did not affect rate but stabilized the β-lactamase; by itself it had no effect. Together mutations A and B were highly beneficial because the β-lactamase was faster and maintained its stability, but adding mutations stepwise will most likely miss these types of synergy. Reetz and Sanchis came to similar conclusions in testing the stepwise addition of mutations to increase enantioselectivity[71]. Synergistic effects are important when one of the mutations is a stabilizing mutation and when the mutations are nearby one another. The complication of non-additivity due to stabilizing mutations can be minimized by stabilizing the protein before starting mutagenesis, but the complication of non-additivity due to nearby mutations is the most common one and not easily avoided.

Consequently, the extent of useful changes made during the improvement of a protein has increased drastically in the past decade. In the early 2000s, 1–5 mutations were typical, whereas by 2010, 30–40 amino-acid

Table 1 | Recently developed biocatalytic processes in the pharmaceutical industry

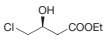
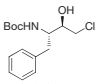
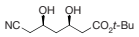
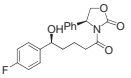
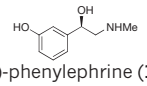
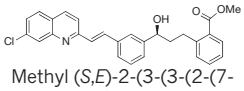
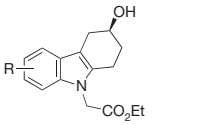
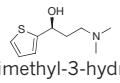
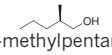
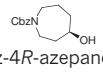
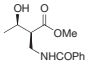
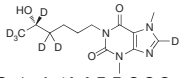
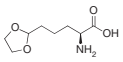
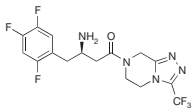
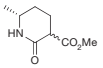
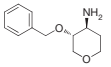
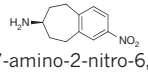
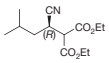
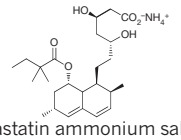
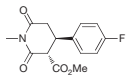
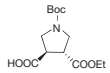
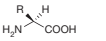
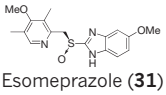
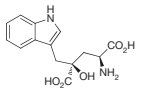
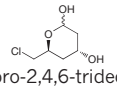
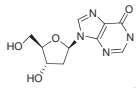
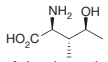
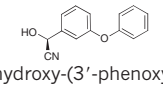
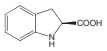
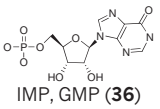
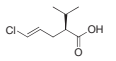
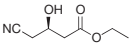
Product	Technology (company, reference)	Product	Technology (company, reference)
KREDs			
 Ethyl (*S*)-4-chloro-3-hydroxybutanoate (**1**)	Various protein engineering approaches, including directed evolution, to increase activity, stability and coenzyme specificity. Initial intermediate for atorvastatin (Lipitor) manufacture on the industrial scale. (Kaneka[81], Codexis[52])	 *Tert*-butyl (2*S*,3*R*)-4-chloro-3-hydroxy-1-phenylbutan-2-yl-carbamate (**10**)	Recombinant expression and also protein engineering for activity. Intermediate for atazanavir (Reyetaz). (Bristol Myers Squibb[82]; Codexis, WO/2011/005527)
 *Tert*-butyl 6-cyano-3,5-dihydroxyhexanoate (**2**)	Protein engineering for activity and stability to 30% organic substrate (the hydroxyketone substrate and a *tert*-butyl acetoacetate impurity, both liquids). Intermediate for atorvastatin. (Codexis, US/2011/7879585; Pfizer, US/2008/0118962)	 (4*S*)-3[(5*S*)-5-(4-fluorophenyl)-5-hydroxy-pentanoyl]-4-phenyl-1,3-oxazolidin-2-one (**11**)	Protein engineering for activity and solvent stability. Intermediate for ezetimibe (Zetia, Vytorin). (Codexis, WO/2010/025085)
 (*R*)-3-hydroxythiolane (R3HT) (**3**)	Directed evolution primarily for enantioselectivity. Intermediate for sulopenem-type antibiotics (since discontinued). (Codexis[83], WO/2009/029554)	 D-pseudoephedrine (**12**)	Protein engineering to overcome product inhibition. The ketone substrate is racemic at C_2 and racemizes during the reaction to give high-e.e. product. The product of the reaction is a generic API. (Daiichi[84])
 (*S*)-1-(2,6-dichloro-3-fluorophenyl)-ethanol (DCFPE) (**4**)	Directed evolution of a variant identified from the R3HT (**3**) project for increased activity. Note that the ketone is hindered owing to the ortho chloro-substituents. Raw material for crizotinib (Xalkori). (Codexis, WO/2009/036404)	 (*R*)-phenylephrine (**13**)	Protein engineering for activity and tolerance to aminoketone substrate. The product of the reaction is a generic API. (Codexis, WO/2011/022548)
 Methyl (*S*,*E*)-2-(3-(3-(2-(7-chloroquinolin-2-yl)vinyl)phenyl)-3-hydroxypropyl) benzoate (**5**)	Directed evolution of a variant identified from the DCFPE (**4**) project. Protein engineering for activity on the poorly soluble substrate (~10 mg l^{-1} in 40% toluene, 10% IPA). Intermediate for montelukast (Singulair). (Codexis[16], WO/2009/042984)	 Ethyl-((7*S*)-7-hydroxy-6,7,8,9-tetrahydro-pyrido[1,2-a]indol-10-yl) acetate (**14**)	A recombinantly expressed, engineered variant was used to produce 1.6 kg of this intermediate for a CRTH2 receptor antagonist. (Merck, US/2010/0234415)
 (*S*)-*N*,*N*-dimethyl-3-hydroxy-3-(2-thienyl)-1-propanamine (**6**)	Protein engineering for activity at pH 9 and IPA, conditions where the ketoamine is stable. Intermediate for duloxetine (Cymbalta). (Codexis, WO/2010/025238)	 (3*R*)-4,4-dimethoxy-tetrahydro-2H-pyran-3-ol (**15**)	Recombinantly produced KRED was used for the stereoselective production of 38 kg of this intermediate. (Merck, US/2008/0138866)
 (*R*)-2-methylpentanol (**7**)	Protein engineering for activity on the aldehyde substrate as well as enantiospecificity towards the desired enantiomer. Intermediate for imagabalin (since discontinued). (Codexis[85], WO/2010/027710)	 *N*-Cbz-4*R*-azepanol (**16**)	A recombinantly expressed enzyme was used to produce this intermediate for a novel β-lactamase inhibitor. (Merck, WO/2008/039420)
 Methyl (2*S*,3*R*)-2-benzamidomethyl-3-hydroxybutyrate (for penem antibiotics) (**8**)	Protein engineering for activity on the racemic ketone as well as for enantiospecificity. The process involves a dynamic kinetic resolution at C_2 to give the desired product in high e.e. and d.e. (Codexis, US/2011/7883879)	 (*R*)-8-d_1-1-(4,4,5,5,6,6,6-d_6-5-hydroxyhexyl)-3,7-dimethyl-1H-purine-2,6-(3H,7H)dione (**17**)	A recombinantly expressed enzyme was used to produce this deuterium containing pentoxyfylline analogue on the gram scale. (Concert Pharmaceuticals, US/2011/0053961)
 (*S*)-Allysine ethylene acetal (for omapatrilat) (**9**)	L-acylase (a hydrolase), large-scale commercial manufacturing. Key intermediate for angiotensin-1-converting enzymes and neutral endopeptidase inhibitors. (Chirotech/DRL[86])	—	—
Transaminases			
 (2*R*)-4-oxo-[3-(trifluoro-methyl)-5,6-dihydro[1,2,3]triazolo[4,3-a]pyrazine-7(8H)-yl]-1-(2,4,5-trifluorophenyl) butan-2-amine (**18**)	Protein engineering to develop enzyme with initial activity, followed by activity improvement in increasing DMSO, substrate and isopropylamine, as well as increased thermostability. The product is sitagliptin (Januvia) free base. (Codexis; Merck[17], WO/2010/099501)	 Methyl (6*R*)-6-methyl-2-oxopiperidine-3-carboxylate (**20**)	A codon-optimized mutant enzyme was used to manufacture 29 kg of this orexin receptor inhibitor. (Merck, WO/2009/143033)
 (3*R*,4*S*)-3-benzyloxy)tetrahydro-2H-pyran-4-amine (**19**)	An enzyme obtained by directed evolution was used for this intermediate for an M1-receptor-positive allosteric regulator. (Merck, WO/2011/062853)	 (7*S*)-7-amino-2-nitro-6,7.8.9-tetrahydro-5H-benzo[7] annulene (**21**)	A recombinantly expressed (codon-optimized) enzyme was used to manufacture this intermediate for an Axl inhibitor on the gram scale. (Rigel, US/2010/0196511)

Table 1 | Continued

Product	Technology (company, reference)	Product	Technology (company, reference)
Hydrolases			
 Intermediate for pregabalin (**22**)	A highly productive process was enabled for this pregabalin intermediate using a natural enzyme after significant reaction engineering. The process produces quantities of the order of 100 tonnes. (Pfizer, WO/2006/000904)	 Simvastatin ammonium salt (**25**)	Reaction engineering and directed evolution enabled a simple process for manufacture of this high-volume API. The enzyme was improved for reaction conditions and a non-natural acyl donor. (UCLA[87]; Codexis, WO/2011/041231)
 Intermediate for paroxetine (**23**)	Intermediates for paroxetine. (BioVerdant, WO/2009/005647)	 L-fluoro-leucine (**26**)	A highly productive process for this odanacatib intermediate was enabled using a natural enzyme after enzyme immobilization and reaction engineering. (Merck[88])
 (1*S*,3*S*)-3-aminocyclohexanol (**24**)	Enzymatic kinetic resolution; lipase from *Thermomyces lanuginosus*; produced on the kilogram scale. (Novartis[89])	 (3*R*,4*R*)-Boc-pyrrolidine-3,4-dicarboxylic acid 3-ethyl ester (**27**)	Cascade reaction using two commercial lipases; >100 kg produced. (Roche[90])
Oxidative enzymes			
 (1*R*,2*S*,5*S*)-6,6-dimethyl-3-azabicyclo[3,1,0] hexane-2-carboxylic acid (**28**)	A fungal amine oxidase was improved for expression in *E. coli* and activity under process conditions, including substrate and product tolerance. The intermediate is used for boceprevir synthesis. (Codexis, WO/2010/008828)	 (*S*)-amino acids (**30**)	(*R*)-amine oxidase/(*S*)-amino acid transferase; key intermediate for antidiabetic drug. (Bristol Myers Squibb[91])
 (1*S*,3a*R*,6a*S*)-octahydrocyclopenta[c]pyrrole-1-carboxylic acid (**29**)	An amine oxidase was improved for activity under process conditions, including substrate and product tolerance. The intermediate is used for telaprevir synthesis. (Codexis, WO/2010/008828)	 Esomeprazole (**31**)	The enantioselectivity of the Baeyer–Villiger monooxygenase was inverted and the enzyme was then improved for activity, stability and chemoselectivity. The product is an API (Nexium). (Codexis, WO/2011/071982)
Aldolases			
 (2*S*,4*S*)-monatin (**32**) (and (2*R*,4*R*)-monatin)	CLEC-subtilisin (chemo-biocatalytic route); kilogram quantities; natural sweetener. (CSIR Biosciences; Altus[92])	 6-chloro-2,4,6-trideoxy-D-erythrohexapyranoside (**33**)	Processes based on 2-deoxy-D-ribose 5-phosphate aldolase were developed by various companies (DSM, Diversa, Pfizer) for this and other statin intermediates. The enzyme was engineered to withstand high concentrations of acetaldehyde and for product selectivity[93]; industrial scale. (DSM, US/2009/0209001)
Other			
 2′-deoxyinosine (**34**)	Deoxyribose aldolase, phosphopentomutase and purine nucleoside phosphorylase were used in a combinatorial biosynthesis. (Yuki Gosei[94])	 (2*S*,3*R*,4*S*)-4-hydroxyleucine (**38**)	L-isoleucine dioxygenase. (Ajinomoto[95])
 2-hydroxy-(3′-phenoxy-2-phenylacetonitrile) (**35**)	Protein engineering of a hydroxynitrile lyase for activity and expression. Used for manufacture of cypermethrin. (DSM, EP/2000/0969095)	 (*S*)-2,3-dihydro-1H-indole-2-carboxylic acid (**39**)	Phenylalanine ammonia lyase; recombinant enzyme from metagenome; industrial scale. (DSM[96])
 IMP, GMP (**36**)	AP/PTase random mutagenesis increased transferase activity and decreased phosphatase activity; industrial scale. (Ajinomoto[97])	 (*S*)-5-chloro-2-isopropylpent-4-enoate (**40**)	Recombinant pig liver esterase; aliskiren intermediate. (DSM[98], WO/2010/10122175)
 D-amino acids (**37**)	D-carbamoylase, thermo- and pH-stability; industrial scale. (Kaneka[99])	 Ethyl (*R*)-4-cyano-3-hydroxybutyrate (**41**)	Directed evolution of a halohydrin dehalogenase for improved activity, stability and tolerance to substrate and product provided a catalyst that is now used for commercial manufacture of this atorvastatin intermediate. (Codexis[52], US/2010/7807423)
Whole cells			
 Artemisinin (**42**)	Pathway engineering; overexpression of additional genes; knockouts; PE. (Amyris; Sanofi-Aventis[100])	 Vanillin (**43**)	Increased productivity; decreased by-product formation. (Givaudan, US/2001/6235507)

API, active pharma ingredient; Boc, butyloxycarbonyl; Cbz, carbobenzyloxy; CLEC, crosslinked enzyme crystal; d.e., diastereomeric excess; DMSO, dimethylsulphoxide; Et, ethyl; GMP, guanosine-5′-monophosphate; IMP, inosine-5′-monophosphate; IPA, isopropyl alcohol; PE, protein engineering.

Table 2 | Presidential Green Chemistry Challenge Awards in biocatalysis over the past ten years

Product	Technology	Company	Year
Succinic acid as chemical feedstock	Fermentation	BioAmber	2011
1,4-butandiol for polymers and chemical feedstock	Fermentation	Genomatica	2011
Higher alcohols as fuels and chemical feedstocks	Fermentation	UCLA (Prof. Dr J. Liao)	2010
Renewable petroleum from fatty-acid metabolism intermediates	Fermentation	LS9	2010
Sitagliptin: a pharmaceutical ingredient for treatment of type 2 diabetes	Enzyme	Merck and Codexis	2010
Esters for cosmetics and personal care products	Enzyme	Eastman Chemical Co.	2009
Atorvastatin intermediate for treatment of high cholesterol	Enzyme	Codexis	2006
Polyhydroxyalkanoates as biodegradable plastics and chemical feedstock	Fermentation	Metabolix	2005
Low trans fats and oils for human nutrition	Enzyme	ADM and Novozymes	2005
Rhamnolipids: biobased, biodegradable industrial surfactants	Fermentation	Jeneil Biosurfactant Company	2004
Taxol for treatment of breast cancer	Fermentation	Bristol Myers Squibb	2004
Improved paper recycling using enzymes to remove sticky contaminants	Enzyme	Buckman Laboratories International	2004
Polyester synthesis using lipases	Enzyme	Polytechnic University (Prof. Dr R. Gross)	2003
1,3-propanediol for polymers	Fermentation	Dupont	2003
Lactic acid for poly(lactic acid) polymers	Fermentation	NatureWorks	2002
Removal of natural waxes and oils from cotton before it is made into fabric	Enzyme	Novozymes	2001

See the US Environmental Protection Agency's website (http://www.epa.gov/greenchemistry/pubs/pgcc/past.html).

substitutions were not unusual. For example, directed evolution of the halohydrin dehalogenase for manufacture of the atorvastatin (Lipitor) side chain (Fig. 3) changed at least 35 of the 254 amino acids[60] (>14%) and directed evolution of the transaminase for sitagliptin manufacture (Fig. 4) changed 27 of the 330 amino acids[17] (8.2%). Similarly, computational design of a retro aldolase required 8 or 12 amino acid substitutions (4–6%) in the starting enzyme, which was a xylanase composed of 197 amino acids[72].

A second approach investigated in the past ten years is the creation of new, often non-natural, catalytic activities. The starting point for this new activity is usually a catalytically promiscuous reaction. Catalytic promiscuity is the ability of one active site to catalyse more than one reaction type. Typically, the enzyme catalyses one normal reaction and additional side reactions, which may involve common catalytic steps. The new reaction type is not just a substituent added to the substrate, but involves a different transition state and/or forms different types of chemical bond. For example, pyruvate decarboxylase normally converts pyruvate to acetaldehyde and carbon dioxide. However, a promiscuous catalytic activity of pyruvate decarboxylase is the coupling of this acetaldehyde to another aldehyde in an acyloin condensation. Such a non-natural pyruvate-decarboxylase-catalysed condensation of acetaldehyde with benzaldehyde is the basis for an industrial process developed at BASF in the 1920s to make a precursor of the drug Ephedrine. Recent protein engineering enhanced the promiscuous ability of pyruvate decarboxylase to catalyse the acyloin condensation[73]. The normal reaction requires a proton transfer, but the promiscuous reaction does not. A single amino-acid substitution to remove the proton donor disabled the natural activity and increased the promiscuous activity about fivefold.

The method of disabling unwanted pathways to increase flux to the desired product is further developed by the third advance, metabolic pathway engineering. This allows more complex pathways from secondary metabolism to be transferred into new organisms and entirely new biochemical pathways to be created to make pharmaceutical intermediates and biofuels. The normal metabolisms of terpenes, amino acids and fatty acids have been re-engineered to make hydrocarbons, alcohols and polyesters for use as fuels, bulk chemicals and plastics (see above).

Challenges remaining in biocatalyst engineering

Despite the advances, there remain major challenges to harnessing the advantages of biocatalysis fully. Enzyme engineering is much faster than it was ten years ago, but changing 30–40 amino acids and screening tens of thousands of candidates still requires a large research team. Many, if not all, engineering strategies will yield improved variants, but some will yield better variants and find them faster. Which ones are the better strategies is still unclear. Directly comparing strategies for the same problem and testing the assumptions behind different strategies will identify the most efficient ones.

The first assumption is that the goal can be achieved using enzyme engineering. The thermodynamics of reactions involving non-natural substrates may be less favourable than that of reactions involving natural substrates, and attaining certain enzyme activities may be thermodynamically impossible. Diffusion sets an upper limit to reaction rates. A closer integration of thermodynamics and biocatalytic process development is highly desirable in designing new processes.

Protein engineering often relies on knowing the quaternary structure of the enzyme because residues at the protein–protein interface can contribute to stability. Researchers assume that the structure of the enzyme under reaction conditions (low enzyme concentrations, high substrate concentration, organic solvents and so on) is very similar to that of the crystallized enzyme (high enzyme concentration, no substrate and/or organic solvent). Because proteins crystallize only under narrow conditions found by extensive experimentation, in solution they probably adopt many conformations besides the ones seen in the crystal structures. Furthermore, our understanding of protein dynamics is still very limited and this makes predictions difficult.

Third, enzyme engineering assumes that individual mutations are additive[67]. Although mutations are mostly non-interactive, many interactive mutations are highly useful but difficult to study. One way of identifying cooperative effects involves statistical analysis using the ProSAR algorithm[60], but better techniques are needed to predict at an early stage of protein engineering which additional mutations are possibly additive and which lead to a dead end.

Fourth, computer design of new enzyme activities is not accurate. Design still requires testing 10–20 predictions and usually results in an enzyme with low activity, which then requires substantial further engineering. For example, the initial computer-based design of an enzyme for the manufacture of sitagliptin[17] yielded an enzyme that converted only 0.1 substrate molecule per day, yet that substrate fits well in the active site within the computer model derived from the crystal structure. New enzymes can be designed to catalyse reactions not found in nature (Kemp-elimination[28], new Diels–Alder reactions[74]), but such activities are so far too low for practical use. Better understanding of the mechanistic, dynamic and structural aspects of enzymatic catalysis is needed.

Technical challenges also limit biocatalysis. The current DNA synthesis methods are close to their efficiency limits, but still cost approximately US$0.35 per base (~US$300 per 1,000-nucleotide gene), which is too high for large-scale applications requiring thousands of genes. Longer and cheaper DNA fragments would simplify and speed up experiments. Next-generation DNA synthesis methods may involve synthesis of oligodeoxynucleotides by codons (trinucleotides) rather than individual nucleotides. This approach was first suggested two decades ago but never reached the mainstream, presumably owing to instrumental limitations (synthesis starts with 64 phosphoramidite trinucleotides). Nevertheless, the concept recently was used in the rapid assembly of entire genes in a single synthesis and in the preparation of high-quality mutagenesis libraries[75], and thus seems feasible today.

New ideas for the integration of biocatalysts with nanodevices and in complex multi-enzyme assemblies hold promise for the future. Enzyme immobilization has been a strategy since the early days of biocatalysis, but it

may be more effective when the biocatalyst's surface orientation is controlled[76]. Similarly, using proteins and nucleic-acid scaffolds to control the number and orientation of enzymes within multi-enzyme pathways also improves efficiency[77]. Separately, functional matrices such as carbon nanotubes and quantum dots can substitute for complex biological electron transfer systems, offering new methods for regenerating redox catalysts and interfacing enzymes with semiconductors[78]. Nevertheless, the integration of enzymes with non-biological matrices and nanomaterials, and as part of metabolic engineering, is still inefficient. Future protein engineering has to address challenges emerging through the interfacing of individual biocatalysts with other proteins in a metabolic pathway or support matrices.

Protein engineering solves the previous weaknesses of biocatalysts: low stability and low activity towards unusual substrates. Large amounts of protein were used to compensate for low activity and this caused emulsions that hampered work-up and reduced yield. Highly active enzymes solve this problem because emulsions do not form using smaller amounts of protein. Training chemists in both biocatalysis and chemocatalysis will help them choose the best solution in each case. Improved enzymes with a long shelf life and good activity and stability in organic solvents should help biocatalysis to spread further into industrial laboratories.

Recent advances in protein engineering have achieved the equivalent of converting mouse proteins into human proteins. The amino-acid sequences of similar proteins in mice and human typically differ by ~13% (ref. 79). Today's advanced protein engineering makes similar changes in converting a wild-type enzyme into an enzyme suitable for chemical process applications. This protein engineering is equivalent to compressing the 75,000,000-yr evolution of an early mammal into modern-day mice and humans into several months of laboratory work. Consistent with the more extensive changes made in these proteins, the properties have also changed more dramatically. The catalytic properties of the enzymes have improved quantitatively by factors of thousands to millions[80], and the engineered enzymes now can act in unusually harsh conditions. The understanding of protein engineering built over the third wave of biocatalysis allows dramatic improvements in enzymatic performance to be realized in parallel with the development of chemical syntheses requiring these catalysts, allowing biocatalysis to develop as an increasingly important tool in chemical synthesis.

1. Buchholz, K., Kasche, V. & Bornscheuer, U. T. *Biocatalysts and Enzyme Technology* 2nd edn (Wiley-VCH, 2012).
2. Drauz, K., Gröger, H. & May, O. (eds) *Enzyme Catalysis in Organic Synthesis* Vols 1–3, 3rd edn (Wiley-VCH, 2012).
3. Bornscheuer, U. T. & Kazlauskas, R. J. *Hydrolases in Organic Synthesis - Regio- and Stereoselective Biotransformations* 2nd edn (Wiley-VCH, 2006).
4. Liese, A., Seelbach, K. & Wandrey, C. (eds) *Industrial Biotransformations* 2nd edn (Wiley-VCH, 2006).
5. Wenda, S., Illner, S., Mell, A. & Kragl, U. Industrial biotechnology—the future of green chemistry? *Green Chem.* **13,** 3007–3047 (2011).
6. Rosenthaler, L. Durch Enzyme bewirkte asymmetrische Synthese. *Biochem. Z.* **14,** 238–253 (1908).
7. Sedlaczek, L. Biotransformations of steroids. *Crit. Rev. Biotechnol.* **7,** 187–236 (1988).
8. Estell, D. A., Graycar, T. P. & Wells, J. A. Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.* **260,** 6518–6521 (1985).
   This paper puts forward the basis for the first application of protein engineering in industrial biotechnology.
9. Jensen, V. J. & Rugh, S. Industrial scale production and application of immobilized glucose isomerase. *Methods Enzymol.* **136,** 356–370 (1987).
10. Bruggink, A., Roos, E. C. & de Vroom, E. Penicillin acylase in the industrial production of β-lactam antibiotics. *Org. Process Res. Dev.* **2,** 128–133 (1998).
11. Griengl, H., Schwab, H. & Fechter, M. The synthesis of chiral cyanohydrins by oxynitrilases. *Trends Biotechnol.* **18,** 252–256 (2000).
12. Hills, G. Industrial use of lipases to produce fatty acid esters. *Eur. J. Lipid Sci. Technol.* **105,** 601–607 (2003).
13. Nagasawa, T., Nakamura, T. & Yamada, H. Production of acrylic acid and methacrylic acid using *Rhodococcus rhodochrous* J1 nitrilase. *Appl. Microbiol. Biotechnol.* **34,** 322–324 (1990).
14. Francis, J. C. & Hansche, P. E. Directed evolution of metabolic pathways in microbial populations. I. Modification of the acid phosphatase pH optimum in *S. cerevisiae*. *Genetics* **70,** 59–73 (1972).
15. Nakamura, C. E. & Whited, G. M. Metabolic engineering for the microbial production of 1,3-propanediol. *Curr. Opin. Biotechnol.* **14,** 454–459 (2003).
16. Liang, J. *et al.* Development of a biocatalytic process as an alternative to the (-)-DIP-Cl-mediated asymmetric reduction of a key intermediate of montelukast. *Org. Process Res. Dev.* **14,** 193–198 (2010).
17. Savile, C. K. *et al.* Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* **329,** 305–309 (2010).
18. Desai, A. A. Sitagliptin manufacture: a compelling tale of green chemistry, process intensification, and industrial asymmetric catalysis. *Angew. Chem. Int. Ed.* **50,** 1974–1976 (2011).
    This highlight article compares in detail two processes for the production of sitagliptin, one catalysed by rhodium and one catalysed by a transaminase.
19. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nature Chem. Biol.* **4,** 617–623 (2008).
20. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451,** 86–89 (2008).
    This paper describes the efficient diversion of amino-acid metabolism to the production of alcohols.
21. Lee, S. K., Chou, H., Ham, T. S., Lee, T. S. & Keasling, J. D. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr. Opin. Biotechnol.* **19,** 556–563 (2008).
22. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463,** 559–562 (2010).
23. Bohmert-Tatarev, K., McAvoy, S., Daughtry, S., Peoples, O. P. & Snell, K. D. High levels of bioplastic are produced in fertile transplastomic tobacco plants engineered with a synthetic operon for the production of polyhydroxybutyrate. *Plant Physiol.* **155,** 1690–1708 (2011).
24. McKenna, R. & Nielsen, D. R. Styrene biosynthesis from glucose by engineered *E. coli*. *Metab. Eng.* **13,** 544–554 (2011).
25. Kazlauskas, R. J. & Bornscheuer, U. T. Finding better protein engineering strategies. *Nature Chem. Biol.* **5,** 526–529 (2009).
26. Lutz, S. & Bornscheuer, U. T. (eds) *Protein Engineering Handbook* (Wiley-VCH, 2009).
27. Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nature Chem. Biol.* **5,** 567–573 (2009).
28. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453,** 190–195 (2008).
29. Schmid, A. *et al.* Industrial biocatalysis today and tomorrow. *Nature* **409,** 258–268 (2001).
30. Arnold, F. H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409,** 253–257 (2001).
31. Schoemaker, H. E., Mink, D. & Wubbolts, M. G. Dispelling the myths—biocatalysis in industrial synthesis. *Science* **299,** 1694–1697 (2003).
32. Keasling, J. D. Manufacturing molecules through metabolic engineering. *Science* **330,** 1355–1358 (2010).
33. Madison, L. L. & Huisman, G. W. Metabolic engineering of poly(3-hydroxyalkanoates): from DNA to plastic. *Microbiol. Mol. Biol. Rev.* **63,** 21–53 (1999).
34. Wetterstrand, K. A. DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. *National Human Genome Research Project* ⟨http://www.genome.gov/sequencingcosts⟩ (2011).
35. Lorenz, P. & Eck, J. Metagenomics and industrial applications. *Nature Rev. Microbiol.* **3,** 510–516 (2005).
36. Fowler, D. M. *et al.* High-resolution mapping of protein sequence- function relationships. *Nature Methods* **7,** 741–746 (2010).
37. Richmond, K. E. *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* **32,** 5011–5018 (2004).
38. Gibson, D. G. *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329,** 52–56 (2010).
39. Lutz, S. Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21,** 734–743 (2010).
40. Höhne, M., Schätzle, S., Jochens, H., Robins, K. & Bornscheuer, U. T. Rational assignment of key motifs for function guides in silico enzyme identification. *Nature Chem. Biol.* **6,** 807–813 (2010).
    In this paper, careful analysis of key motifs of 5,000 pyridoxal-5-phosphate-dependent transaminase sequences in public databases identified 20 novel enzymes for which substrate preference (ketone, not α-keto acid) and enantiopreference ((*R*), not (*S*)) could be predicted and experimentally confirmed.
41. Jochens, H. & Bornscheuer, U. T. Natural diversity to guide focused directed evolution. *ChemBioChem* **11,** 1861–1866 (2010).
42. Park, S. *et al.* Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem. Biol.* **12,** 45–54 (2005).
43. Horsman, G. P., Liu, A. M. F., Henke, E., Bornscheuer, U. T. & Kazlauskas, R. J. Mutations in distant residues moderately increase the enantioselectivity of *Pseudomonas fluorescens* esterase towards methyl 3-bromo-2-methylpropanoate and ethyl 3-phenylbutyrate. *Chemistry* **9,** 1933–1939 (2003).
44. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472,** 499–503 (2011).
45. Moore, J. C., Pollard, D. J., Kosjek, B. & Devine, P. N. Advances in the enzymatic reduction of ketones. *Acc. Chem. Res.* **40,** 1412–1419 (2007).
46. Strohmeier, G. A., Pichler, H., May, O. & Gruber-Khadjawi, M. Application of designed enzymes in organic synthesis. *Chem. Rev.* **111,** 4141–4164 (2011).
    This is a review about protein engineering to create biocatalysts for industrial applications.
47. Wiese, A., Wilms, B., Syldatk, C., Mattes, R. & Altenbuchner, J. Cloning, nucleotide sequence and expression of a hydantoinase and carbamoylase gene from

Arthrobacter aurescens DSM 3745 in *Escherichia coli* and comparison with the corresponding genes from *Arthrobacter aurescens* DSM 3747. *Appl. Microbiol. Biotechnol.* **55,** 750–757 (2001).
48. Martinez-Gomez, A. I. *et al.* Recombinant polycistronic structure of hydantoinase process genes in *Escherichia coli* for the production of optically pure D-amino acids. *Appl. Environ. Microbiol.* **73,** 1525–1531 (2007).
49. Panke, S. & Wubbolts, M. Advances in biocatalytic synthesis of pharmaceutical intermediates. *Curr. Opin. Chem. Biol.* **9,** 188–194 (2005).
50. Breuer, M. *et al.* Industrial methods for the production of optically active intermediates. *Angew. Chem. Int. Ed.* **43,** 788–824 (2004).
51. DeSantis, G. *et al.* Creation of a productive, highly enantioselective nitrilase through gene site saturation mutagenesis (GSSM). *J. Am. Chem. Soc.* **125,** 11476–11477 (2003).
This paper describes how a single amino-acid substitution can create a nitrilase with high enantioselectivity at 3 M substrate concentration, for synthesis of an intermediate for atorvastatin.
52. Ma, S. K. *et al.* A green-by-design biocatalytic process for atorvastatin intermediate. *Green Chem.* **12,** 81–86 (2010).
53. Anastas, P. & Warner, J. (eds) *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
54. Yang, T. H. *et al.* Biosynthesis of polylactic acid and its copolymers using evolved propionate CoA transferase and PHA synthase. *Biotechnol. Bioeng.* **105,** 150–160 (2010).
55. Bernath, K. *et al. In vitro* compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal. Biochem.* **325,** 151–157 (2004).
56. Becker, S. *et al.* Single-cell high-throughput screening to identify enantioselective hydrolytic enzymes. *Angew. Chem. Int. Ed.* **47,** 5085–5088 (2008).
57. Fernández-Alvaro, E. *et al.* A combination of *in vivo* selection and cell sorting for the identification of enantioselective biocatalysts. *Angew. Chem. Int. Ed.* **50,** 8584–8587 (2011).
58. Whittle, E. & Shanklin, J. Engineering delta 9–16:0-acyl carrier protein (ACP) desaturase specificity based on combinatorial saturation mutagenesis and logical redesign of the castor delta 9–18:0-ACP desaturase. *J. Biol. Chem.* **276,** 21500–21505 (2001).
59. Seelig, B. & Szostak, J. W. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* **448,** 828–831 (2007).
60. Fox, R. J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nature Biotechnol.* **25,** 338–344 (2007).
61. Yoshikuni, Y., Ferrin, T. E. & Keasling, J. D. Designed divergent evolution of enzyme function. *Nature* **440,** 1078–1082 (2006).
62. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., &. Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* **32,** D120–D121 (2004).
63. Guo, H. H., Choe, J. & Loeb, L. A. Protein tolerance to random amino acid change. *Proc. Natl Acad. Sci. USA* **101,** 9205–9210 (2004).
This paper shows that about 34% of random amino-acid replacements inactivate a protein's functions, indicating the importance of starting with a stabilized protein for protein engineering experiments.
64. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA* **102,** 14338–14343 (2005).
65. Gupta, R. D. & Tawfik, D. S. Directed enzyme evolution via small and effective neutral drift libraries. *Nature Methods* **5,** 939–942 (2008).
66. Bloom, J. D., Romero, P. A., Lu, Z. & Arnold, F. H. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biol. Direct* **2,** 17 (2007).
67. Wells, J. A. Additivity of mutational effects in proteins. *Biochemistry* **29,** 8509–8517 (1990).
68. Moore, J. C. & Arnold, F. H. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature Biotechnol.* **14,** 458–467 (1996).
69. Reetz, M. T., Zonta, A., Schimossek, K., Liebeton, K. & Jaeger, K.-E. Creation of enantioselective biocatalysts for organic chemistry by *in vitro* evolution. *Angew. Chem. Int. Edn Engl.* **36,** 2830–2832 (1997).
70. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).
71. Reetz, M. T. & Sanchis, J. Constructing and analyzing the fitness landscape of an experimental evolutionary process. *ChemBioChem* **9,** 2260–2267 (2008).
72. Jiang, L. *et al. De novo* computational design of retro-aldol enzymes. *Science* **319,** 1387–1391 (2008).
73. Meyer, D. *et al.* Conversion of pyruvate decarboxylase into an enantioselective carboligase with biosynthetic potential. *J. Am. Chem. Soc.* **133,** 3609–3616 (2011).
74. Siegel, J. B. *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* **329,** 309–313 (2010).
75. Randolph, J., Yagodkin, A., Lamaitre, M., Azhayev, A. & Mackie, H. Codon based mutagenesis using trimer phosphoramidites. *Nucleic Acids Symp. Ser.* **52,** 479 (2008).
76. Rana, S., Yeh, Y. C. & Rotello, V. M. Engineering the nanoparticle-protein interface: applications and possibilities. *Curr. Opin. Chem. Biol.* **14,** 828–834 (2010).
77. Dueber, J. E. *et al.* Synthetic protein scaffolds provide modular control over metabolic flux. *Nature Biotechnol.* **27,** 753–759 (2009).
78. McDonald, T. J. *et al.* Wiring-up hydrogenase with single-walled carbon nanotubes. *Nano Lett.* **7,** 3528–3534 (2007).
79. Mouse Genome Sequencing Consortium.. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
80. Savile, C. K. & Lalonde, J. J. Biotechnology for the acceleration of carbon dioxide capture and sequestration. *Curr. Opin. Biotechnol.* **22,** 818–823 (2011).
81. Morikawa, S. *et al.* Highly active mutants of carbonyl reductase S1 with inverted coenzyme specificity and production of optically active alcohols. *Biosci. Biotechnol. Biochem.* **69,** 544–552 (2005).
82. Patel, R. N., Chu, L. & Mueller, R. Diastereoselective microbial reduction of (S)-[3-chloro-2-oxo-1-(phenylmethyl)propyl]carbamic acid, 1,1-dimethylethyl ester. *Tetrahedr. Asymm.* **14,** 3105–3109 (2003).
83. Liang, J. *et al.* Highly enantioselective reduction of a small heterocyclic ketone: biocatalytic reduction of tetrahydrothiophene-3-one to the corresponding (*R*)-alcohol. *Org. Process Res. Dev.* **14,** 188–192 (2010).
84. Urano, N. *et al.* Directed evolution of an aminoalcohol dehydrogenase for efficient production of double chiral aminoalcohols. *J. Biosci. Bioeng.* **111,** 266–271 (2011).
85. Gooding, O. W. *et al.* Development of a practical biocatalytic process for (*R*)-2-methylpentanol. *Org. Process Res. Dev.* **14,** 119–126 (2010).
86. Cobley, C. J., Hanson, C. H., Lloyd, M. C. & Simmonds, S. The combination of hydroformylation and biocatalysis for the large-scale synthesis of (*S*)-allysine ethylene acetal. *Org. Process Res. Dev.* **15,** 284–290 (2011).
87. Xie, X., Watanabe, K., Wojcicki, W. A., Wang, C. C. & Tang, Y. Biosynthesis of lovastatin analogs with a broadly specific acyltransferase. *Chem. Biol.* **13,** 1161–1169 (2006).
88. Truppo, M. D. & Hughes, G. Development of an improved immobilized CAL-B for the enzymatic resolution of a key intermediate to odanacatib. *Org. Process Res. Dev.* **15,** 1033–1035 (2011).
89. Brocklehurst, C. E., Laumen, K., Vecchia, L. L., Shaw, D. & Vögtle, M. Diastereoisomeric salt formation and enzyme-catalyzed kinetic resolution as complementary methods for the chiral separation of cis-/trans-enantiomers of 3-aminocyclohexanol. *Org. Process Res. Dev.* **15,** 294–300 (2011).
90. Iding, H. *et al.* in *Asymmetric Catalysis on Industrial Scale* (eds Blaser, H. U. & Federsel, H. J.) 377–396 (Wiley-VCH, 2010).
91. Chen, Y. J. *et al.* Enzymatic preparation of an (*S*)-amino acid from a racemic amino acid. *Org. Process Res. Dev.* **15,** 241–248 (2011).
92. Rousseau, A. L. *et al.* Scale-up of a chemo-biocatalytic route to (2*R*,4*R*)- and (2*S*,4*S*)-monatin. *Org. Process Res. Dev.* **15,** 249–257 (2011).
93. Greenberg, W. A. *et al.* Development of an efficient, scalable, aldolase-catalyzed process for enantioselective synthesis of statin intermediates. *Proc. Natl Acad. Sci. USA* **101,** 5788–5793 (2004).
94. Horinouchi, N. *et al.* Biochemical retrosynthesis of 2′-deoxyribonucleosides from glucose, acetaldehyde, and a nucleobase. *Appl. Microbiol. Biotechnol.* **71,** 615–621 (2006).
95. Hibi, M. *et al.* Characterization of *Bacillus thuringiensis L*-isoleucine dioxygenase for production of useful amino acids. *Appl. Environ. Microbiol.* **77,** 6926–6930 (2011).
96. de Lange, B. *et al.* Asymmetric synthesis of (*S*)-2-indoline carboxylic acid by combining biocatalysis and homogeneous catalysis. *ChemCatChem* **3,** 289–292 (2011).
97. Asano, Y., Mihara, Y. & Yamada, H. A new enzymatic method of selective phosphorylation of nucleosides. *J. Mol. Catal. B* **6,** 271–277 (1999).
98. Hermann, M. *et al.* Alternative pig liver esterase (APLE) - cloning, identification and functional expression in *Pichia pastoris* of a versatile new biocatalyst. *J. Biotechnol.* **133,** 301–310 (2008).
99. Ikenaka, Y. *et al.* Thermostability reinforcement through a combination of thermostability-related mutations of N-carbamyl-D-amino acid amidohydrolase. *Biosci. Biotechnol. Biochem.* **63,** 91–95 (1999).
100. Ro, D. K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 940–943 (2006).

Acknowledgements We thank H.-P. Meyer and R. Fox for discussions. R.J.K. thanks the US National Science Foundation (CBET-0932762) and the Korea Science and Engineering Foundation funded by the Ministry of Education, Science and Technology (WCU programme R32-2008-000-10213-0). U.T.B. and S.L. thank the German Research Foundation (SPP 1170, Bo1864/4-1) and, respectively, the US National Science Foundation (CBET-0730312) for financial support.

Author Contributions U.T.B., R.J.K. and S.L. drafted the text; K.R., G.W.H. and J.C.M. collected the examples for industrial applications; and the authors together wrote and edited the review. All authors contributed equally.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to U.T.B. (uwe.bornscheuer@uni-greifswald.de).

BLACK HOLES

Star ripped to shreds

When a star wanders too close to a giant black hole, it can be pulled apart by the black hole's tidal force. One such event offers insight into the properties of both the black hole and the star. SEE LETTER P.217

GIUSEPPE LODATO

Astronomers have strong evidence that supermassive black holes, with masses between a million and a billion times that of the Sun, reside in the centre of most galaxies. The evidence comes from observations of the copious amount of radiation that is emitted when these objects pull gas from their immediate vicinity. However, if a black hole's close environment is poor in gas, gas accretion proceeds at a low rate and is not accompanied by significant emission of radiation. Probing such 'dormant' black holes is therefore difficult — unless a tidal-disruption event occurs. Such an event takes place when a star comes close enough to the black hole to be ripped apart by its tidal force. The ensuing stellar debris is accreted by the black hole and produces a characteristic flare. On page 217 of this issue, Gezari *et al.*[1] describe how detailed observations of a tidal-disruption event have allowed them to determine the properties of not only the black hole, but also the disrupted star*.

Tidal-disruption events are rare. They are expected to occur once every 10,000 years per galaxy. To find them, it is therefore necessary to use large astronomical surveys in which thousands of galaxies are regularly observed. Gezari and colleagues' discovery comes from one such survey, which is expected to detect roughly one event every two years[2].

The main significance of this study probably lies in the accurate determination of the properties of the tidal-disruption event, which the authors based on a well-sampled ultraviolet–optical light curve for the flare (a plot that shows the evolution of the flare's brightness over time) and spectroscopic measurements of the system. They find that the supermassive black hole is hosted by a galaxy at redshift 0.17 (corresponding to a distance of approximately 2 billion light years from Earth), and has a mass of about 3 million solar masses. Moreover, on the basis of the shape of the light curve[3] and the absence of hydrogen lines in the spectra of the stellar debris, Gezari *et al.* conclude that the disrupted star was the helium-rich core of a red giant whose hydrogen outer shell had been previously stripped off, possibly by the same tidal force that eventually led to its complete disruption (Fig. 1).

Gezari and colleagues' observations also imply that the orbit of the star around the black hole was exceptionally tight, with the point of closest approach being only six times the black hole's Schwarzschild radius, which, for a non-rotating black hole, corresponds to its event horizon — the boundary beyond which nothing, not even light, can escape. After the point of closest approach, part of the debris was expelled from the system and part was launched into highly eccentric orbits, falling onto the black hole after approximately two months from closest approach, and producing the observed ultraviolet–optical flare.

During the past year, two other tidal-disruption events caused by supermassive black holes have been detected[4–7]. In those cases, the emission occurred over a wide range of the electromagnetic spectrum, from X-rays to radio waves, because it was produced by a high-energy jet of particles that happened to point almost exactly in our line of sight; such jets are expected to be associated with tidal-disruption events. By contrast, the ultraviolet–optical radiation of tidal-disruption events is usually associated with thermal emission from an accretion disk of stellar debris that forms around the black hole. However, somewhat surprisingly, in the present case Gezari and colleagues find that the ultraviolet–optical emission does not seem to be caused by an accretion disk, because its emission would have fallen off with time[8] at a different rate from that observed. This finding will provide food for thought for subsequent investigations of these events.

Finally, the observation that the star's closest approach to the black hole corresponds to a distance of only six times the Schwarzschild radius suggests that effects of the general theory of relativity might need to be invoked to describe the system. Would the shape of the light curve be modified by such effects? Current models, including those used by Gezari *et al.*, are generally based on Newtonian dynamics, and hence do not include the effects of general relativity. As a result, the models cannot distinguish between a rotating and a non-rotating black hole. Tidal-disruption events can probe deeply into the gravitational field of the black hole, and so offer the means to test the effects of general relativity and black-hole rotation. The fact that we are now in a position to characterize the light curve and spectra of such events so accurately, and that the rate of discovery seems to be in line with expectations, means that we should soon be able to answer these fascinating questions. ■

*This article and the paper[1] under discussion were published online on 2 May 2012.

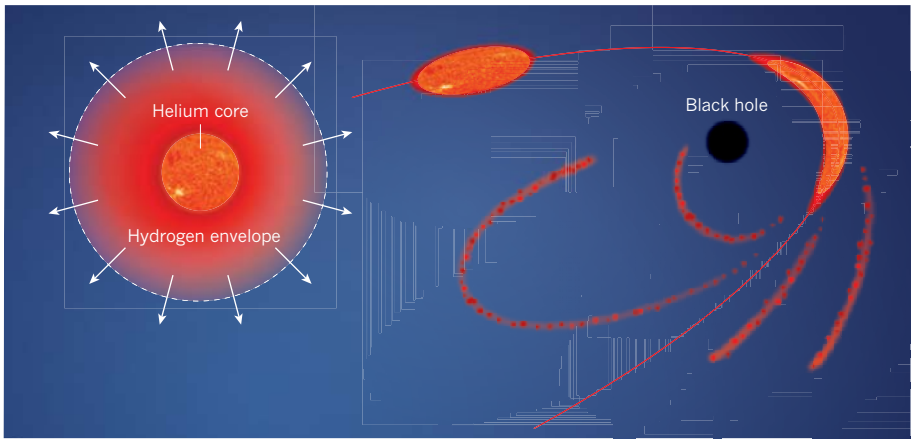


Figure 1 | Tidal disruption of a star. The helium-rich core of a red-giant star that had previously lost its hydrogen envelope moves on an almost parabolic orbit (red) towards a supermassive black hole. The sequence of blobs illustrates the progressive distortion of the star's core due to the tidal pull of the black hole. After the point of closest approach to the black hole, the core is completely disrupted, with part of the resulting debris being expelled from the system and part being launched into highly eccentric orbits, eventually falling onto the black hole. Accretion of this debris gives rise to the intense ultraviolet–optical flare that has been observed by Gezari and colleagues[1].

Giuseppe Lodato *is in the Physics Department, University of Milan, I-20133 Milan, Italy.*
e-mail: giuseppe.lodato@unimi.it

1. Gezari, S. *et al. Nature* **485,** 217–220 (2012).
2. Gezari, S. *et al. Astro2010: The Astronomy and Astrophysics Decadal Survey* Science White Pap. 88; preprint at http://arxiv.org/abs/0903.1107 (2009).
3. Lodato, G., King, A. R. & Pringle, J. E. *Mon. Not. R. Astron. Soc.* **392,** 332–340 (2009).
4. Bloom, J. S. *et al. Science* **333,** 203–206 (2011).
5. Burrows, D. N. *et al. Nature* **476,** 421–424 (2011).
6. Zauderer, B. A. *et al. Nature* **476,** 425–428 (2011).
7. Cenko, S. B. *et al.* preprint at http://arxiv.org/abs/1107.5307 (2011).
8. Lodato, G. & Rossi, E. M. *Mon. Not. R. Astron. Soc.* **410,** 359–367 (2011).

ARTICLE

doi:10.1038/nature11112

Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq

Dan Dominissini[1,2]*, Sharon Moshitch-Moshkovitz[1]*, Schraga Schwartz[3]*†, Mali Salmon-Divon[1], Lior Ungar[2,4], Sivan Osenberg[1,2], Karen Cesarkas[1], Jasmine Jacob-Hirsch[1], Ninette Amariglio[1], Martin Kupiec[4], Rotem Sorek[3] & Gideon Rechavi[1,2]

An extensive repertoire of modifications is known to underlie the versatile coding, structural and catalytic functions of RNA, but it remains largely uncharted territory. Although biochemical studies indicate that N^6-methyladenosine (m^{6}A) is the most prevalent internal modification in messenger RNA, an in-depth study of its distribution and functions has been impeded by a lack of robust analytical methods. Here we present the human and mouse m^{6}A modification landscape in a transcriptome-wide manner, using a novel approach, m^{6}A-seq, based on antibody-mediated capture and massively parallel sequencing. We identify over 12,000 m^{6}A sites characterized by a typical consensus in the transcripts of more than 7,000 human genes. Sites preferentially appear in two distinct landmarks—around stop codons and within long internal exons—and are highly conserved between human and mouse. Although most sites are well preserved across normal and cancerous tissues and in response to various stimuli, a subset of stimulus-dependent, dynamically modulated sites is identified. Silencing the m^{6}A methyltransferase significantly affects gene expression and alternative splicing patterns, resulting in modulation of the *p53* (also known as *TP53*) signalling pathway and apoptosis. Our findings therefore suggest that RNA decoration by m^{6}A has a fundamental role in regulation of gene expression.

Although the multiple layers of epigenetic regulation that result from modification of DNA and proteins have been intensively explored, RNA modifications are still uncharted territory[1]. The complex structure–function relationship of RNA makes it challenging to decipher its intricate biological roles. A wide variety of post-transcriptional modifications, with over a hundred known so far[2], decorate RNA molecules from all domains of life to expand their nucleotide repertoire. Nonetheless, knowledge of their location and function is limited at present.

m^{6}A is the most common internal messenger RNA modification found in eukaryotes, as well as in RNA of nuclear-replicating viruses[3]. It is catalysed by an evolutionarily conserved, nuclear, multi-component enzyme, only one of whose subunits, methyltransferase like 3 (*METTL3*), has been identified[4]. In all organisms tested, induced experimental deficiency of the methyltransferase is detrimental and leads to apoptosis (*Homo sapiens*[3]), developmental arrest (*Arabidopsis thaliana*[5]) or defects in gametogenesis (*Saccharomyces cerevisiae*[6], *Drosophila melanogaster*[7]).

Since m^{6}A was first discovered in mRNA decades ago[8], only a few sites have been mapped in cellular and viral RNA, none of them in human, both in coding and non-coding regions[9,10]. All sites were found within sequences conforming to the degenerate consensus RRACH (A = m^{6}A)[11,12]. Unlike adenosine-to-inosine editing, m^{6}A does not alter the coding capacity of transcripts[13,14]. Importantly, it was shown that a specific position can be methylated in only a fraction of transcripts[9] (that is, non-stoichiometric) and that certain transcripts seem to be completely devoid of m^{6}A (ref. 15), suggesting that this base modification serves some regulatory role. The recent identification of fat mass and obesity-associated protein (*FTO*) as the m^{6}A demethylase[16], and its dynamic association with nuclear speckles further supports this notion.

Progress towards elucidating the biological function of m^{6}A was largely impeded by a lack of efficient methods for its detection and manipulation[13]. Although clearly attesting to the essential role of the m^{6}A methyltransferase, previous studies provided only limited insight into the underlying molecular mechanisms by which this ubiquitous modification acts (reviewed in ref. 3).

To gain further insight into the role of m^{6}A in RNA metabolism, we sought to determine the positions of m^{6}As at a transcriptome-wide level. We present a novel approach, m^{6}A-seq, which accomplishes this goal by combining the high specificity of an anti-m^{6}A antibody applied to randomly fragmented transcripts with the power of massively parallel sequencing. Using m^{6}A-seq, we compile the first human and mouse RNA methylomes, to our knowledge, and demonstrate their evolutionary conservation and response to changing cellular conditions. We complement this analysis by characterizing the effects of m^{6}A on splicing, and by identifying protein candidates associated with methylated RNA sequences (Supplementary Fig. 1).

m^{6}A-seq exposes the RNA methylome

To identify and localize m^{6}A sites at a transcriptome-wide level we applied m^{6}A-seq to RNA purified from a human hepatocellular carcinoma cell line (HepG2). Poly(A)$^+$-selected RNA was fragmented into ~100-nucleotide-long oligonucleotides (input) and immunoprecipitated using an anti-m^{6}A affinity purified antibody. Libraries were prepared from immunoprecipitated as well as input control fragments, and subjected to massively parallel sequencing (Fig. 1a). Reads were uniquely aligned to a reference transcriptome containing a single, intron-free splice variant per gene (Supplementary Table 2), and m^{6}A sites were identified using a peak-detection algorithm with an estimated false detection rate (FDR) < 7% (Methods). This analysis yielded 12,769 putative m^{6}A sites (henceforth referred to as m^{6}A peaks) within 6,990 coding gene transcripts (Fig. 2e and Supplementary Fig. 2a–c) and 250 non-coding ones (Supplementary Fig. 2d),

[1]Cancer Research Center, Chaim Sheba Medical Center, Tel Hashomer 52621, Israel. [2]Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel. [3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel. [4]Department of Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv 69978, Israel. †Present address: Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA.
*These authors contributed equally to this work.

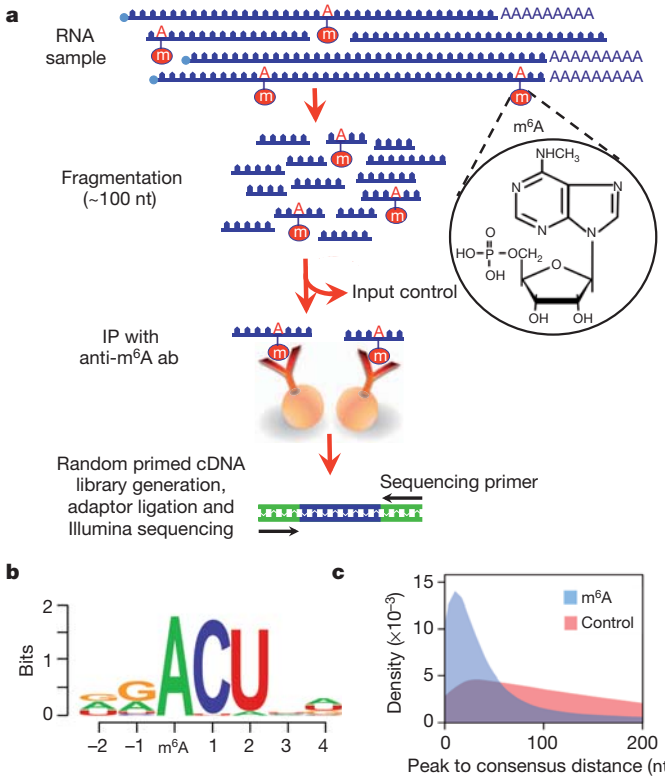


Figure 1 | m^{6}A-seq capture of modified RNA fragments exposes an enriched motif. **a**, Schematic diagram of the m^{6}A-seq protocol. ab, antibody; nt, nucleotide. IP, immunoprecipitation. **b**, Sequence logo representing the deduced consensus motif following clustering of all enriched motifs. **c**, Density plot illustrating the distribution of distance between the peaks of m^{6}A/control fragments and the nearest m^{6}A consensus motif (RRACU).

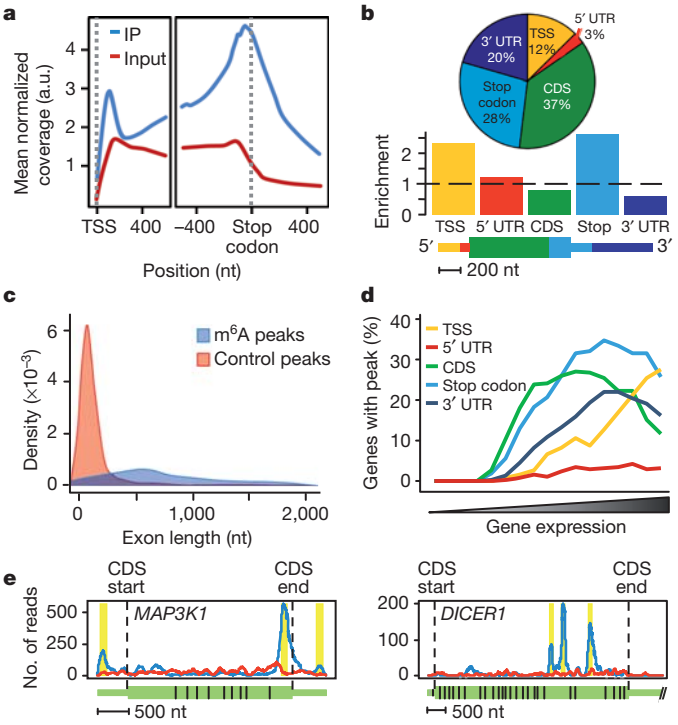


Figure 2 | The transcriptome landscape of m^{6}A reveals a unique topology. **a**, Metagene profiles depicting sequence coverage in windows surrounding the TSS (left) and stop codon (right). Coverage of m^{6}A IP and control (input) fragments is indicated in blue and red, respectively. To normalize for expression levels, coverage for each position is divided by the mean coverage in the control experiment. **b**, Top, pie chart presenting the fraction of m^{6}A peaks in each of five non-overlapping transcript segments. Middle, relative enrichment of m^{6}A peaks across transcript segments. Bottom, schematic of the five segments. **c**, Density plot showing the distribution of m^{6}A/control peaks according to exon length. **d**, Fraction of genes with m^{6}A peaks in each of the segments as a function of expression level. **e**, Representative gene transcripts harbouring m^{6}A peaks. Colour codes are the same as in panel **a**. Peaks are highlighted in yellow. Black dashed lines signify CDS borders; transcript architecture is shown beneath, with thin parts corresponding to UTRs and thicker ones to CDS; exon–exon junctions are indicated by vertical black lines. a.u., arbitrary units.

such as lincRNAs and antisense RNAs—comprising the first m^{6}A RNA methylome (Supplementary Tables 1, 6 and Supplementary Link 1). This number probably underestimates the actual number of m^{6}A sites (Supplementary Fig. 3).

Repeated analysis using the model-based analysis of ChIP-seq (MACS) peak-calling algorithm based on genome-aligned reads identified 20,401 peaks (FDR ≤ 5%, fold change ≥ 4), containing over 96% of the transcriptome-based peaks, among them 1,508 intron peaks (Supplementary Table 3).

The high specificity and immunoprecipitation compatibility of the antibody used are well established and have been described in a multitude of publications[9,16–22]. Several additional measures were taken to ensure the validity and stringency of our experimental approach (Supplementary Note 1), in addition to identification of a known m^{6}A site within 18S rRNA[23].

To determine whether our identified m^{6}A peaks share a common sequence element potentially specifying methylation, we performed an unbiased search for motifs enriched in regions surrounding m^{6}A peaks (Methods). Clustering of all significantly enriched sequences perfectly recapitulated the previously established RRAC core, and revealed a terminal U to be a dominant part of the consensus (Fig. 1b and Supplementary Fig. 4). The median distance between m^{6}A peaks and the consensus sequence was 24 nucleotides, compared to 117 nucleotides in control peaks, supporting the specificity of m^{6}A-seq and reflecting its resolution (Fig. 1c). The identification of a strong consensus reinforces the authenticity of the newly discovered m^{6}A peaks, and supports the existence of a predominant methylation machinery.

Our results show that the HepG2 transcriptome contains ~1 m^{6}A peak per 2,000 nucleotides (Supplementary Fig. 3), or ~1.7 peaks per gene. These results are in line with previous estimations of ~3 m^{6}A residues per average mRNA transcript in mammalian cells, which largely relied on the gross proportion of m^{6}As out of all methylated nucleosides[24]. Of note, as transcripts are fragmented before immunoprecipitation, the m^{6}A-seq analysis is blind to methylation patterns at the single transcript level, but rather produces a gene-specific methylation profile based on fragments derived from the entire transcript population of a given gene.

Given that the determined consensus is far more prevalent than the actual observed frequency of m^{6}As by an order of magnitude, other structural elements, sequence determinants and/or processing steps might be necessary to specify a site for methylation. A previous small-scale computational prediction suggested that m^{6}As lie within the loop of a stem–loop structure[3]. However, using RNAfold, we were unable to identify strong secondary structures in regions harbouring newly identified m^{6}A peaks.

As can be expected from such widespread occurrence, no functional category (GOTERM_MF_1-5) was found to be enriched in a gene ontology (GO) analysis, suggesting that m^{6}A has a fundamental function.

m^{6}A distribution reveals a unique topology

Having exposed a new layer of RNA modification, the next challenge is to infer putative functions. To this end, we began by systematically examining the distribution of m^{6}A peaks along transcripts relative to

landmarks in their architecture, based on the single-transcript, intron-free transcriptome data set (to avoid misinterpretations due to peaks falling into different isoform-dependent landmarks). We noted that m^{6}A peaks were markedly correlated with two distinct coordinates: immediately following the transcription start site (TSS) and in the vicinity of the stop codon (Fig. 2a). Overall, stop codon peaks were more pronounced than TSS peaks. To assess the enrichment methodically, we assigned each m^{6}A peak to one of five non-overlapping transcript segments: TSS; 5′ untranslated region (UTR); coding sequence (CDS); stop codon; and 3′ UTR (Fig. 2b, pie chart), and then normalized by the relative fraction each segment occupied in the transcriptome (Fig. 2b, histogram). The stop codon segment (400-nucleotide window centred on the stop codon) stood out as most enriched in m^{6}A peaks, with 28% of the peaks, representing a ~2.6-fold enrichment over the distribution expected by chance ($P < 2.2 \times 10^{-308}$, χ^2 test; Fig. 2e and Supplementary Fig. 5). Moreover, 33.5% of all adequately expressed genes (that is, >40 reads per kilobase (kb)) had a peak in this region. In only 2.3% of the cases did a methylation consensus form at the stop codon (incorporating its two consecutive purines), indicating that methylation typically did not occur precisely at the stop codon but rather in its vicinity.

Nonetheless, the much longer CDS segment, although slightly depleted in m^{6}A peaks when normalized by length, harboured the largest fraction of peaks (37%) (Fig. 2b, e). Notably, these peaks tend to occur within unusually long internal exons: of 2,838 m^{6}A peaks within internal exons, 2,453 (87%) are in exons longer than 400 nucleotides, compared to only 14% of the negative control peaks ($P = 3.4 \times 10^{-156}$, *t*-test) (Fig. 2c, Supplementary Fig. 6 and Supplementary Table 1). Comparing the lengths of flanking introns and the strengths of the 5′ and 3′ splice sites and polypyrimidine tract, we found only minor differences between long methylated exons and their unmethylated counterparts, which for the most part were not statistically significant.

In contrast to all other segments, we found no tendency for TSS peaks (that is, in the first 200 nucleotides of a transcript) to be near a methylation consensus sequence (Supplementary Fig. 7b), suggesting that a methyltransferase other than the *METTL3* complex is involved. Indeed, when adenosine is the first transcribed nucleotide, in addition to the obligatory ribose 2′-*O*-methylation, it can be further methylated by another methyltransferase at the N^6 position of the base to generate (N^6-,2′-*O*-)-dimethyladenosine (m^{6}A$_m$)[25,26], also recognizable by our antibody[21]. In support of this claim, transcripts with TSS peaks were ~25% more likely to begin with adenosine compared with transcripts lacking such peaks (Supplementary Fig. 7a). Taken together, TSS peaks appear to reflect, at least partly, m^{6}A$_m$ belonging to the 5′ cap structure, and attest, once more, to the validity of m^{6}A-seq.

We next examined the relationship between methylation and expression by plotting the fraction of genes with m^{6}A peaks in each of the segments as a function of expression level. Interestingly, the CDS, stop codon and 3′ UTR segments exhibited a non-monotonic relationship: whereas transcripts of moderately expressed genes were more likely to be methylated, transcripts of genes expressed at the two extremes were less methylated (Fig. 2d). This pattern, which is unlikely to be due to coverage limitations or to bias of differential read sampling (Supplementary Fig. 8), is of functional interest and is reminiscent of a similar relationship between DNA methylation found in gene bodies and their corresponding expression levels[27], raising the intriguing possibility that these phenomena might be connected. In contrast, we observed a positive correlation between expression level and the presence of an m^{6}A peak in the TSS segment.

Comparison of human and mouse methylomes

The non-random distribution of m^{6}A peaks in the human transcriptome indicates a fundamental role of this modification. To determine the evolutionary conservation and consequent functional importance of m^{6}A, the human and mouse methylomes were compared. We applied m^{6}A-seq to RNA purified from mouse liver and obtained 4,513 m^{6}A peaks within 3,376 coding gene transcripts and 66 non-coding ones (Supplementary Tables 1 and 7). Clustering of all significantly enriched sequences perfectly recapitulated the human methylation consensus sequence (Fig. 3a and Supplementary Fig. 9a). The mouse metagene profile revealed, as in human, a peak around the stop codon and at the TSS (Fig. 3b and Supplementary Fig. 9c, d). In mouse, too, modifications were highly enriched in long internal exons, with 91% of the peaks in exons longer than 400 nucleotides (Supplementary Table 1 and Supplementary Fig. 9b). Thus, on the global level, m^{6}A is highly conserved between the two species.

We next systematically assessed the extent of m^{6}A peak conservation on the gene level. Of the 4,513 identified mouse m^{6}A peaks, 2,023 could be reliably mapped to an adequately expressed, orthologous human position; 997 of them had an m^{6}A peak in the orthologous human position, representing 49% conservation. The highest extent of conservation is found at the stop codon segment, in which 57% of peaks are conserved, compared to 32–49% in other segments. Conservation was also higher in internal exons longer than 400 nucleotides than in shorter ones (56% and 44%, respectively).

The extent of conservation between human and mouse, which is highly significant over that which would be expected by chance ($P = 3.7 \times 10^{-136}$, Mann–Whitney test), is easily appreciated in the examples presented (Fig. 3c and Supplementary Fig. 10), and signifies that these sites are likely to be functional.

Methylation across conditions and cells

Accumulating evidence reveals that transfer RNA (tRNA) and ribosomal RNA (rRNA) modifications change in response to stimuli, suggesting a general model of dynamic control over RNA modification[28,29]. By analogy, we reasoned that participation of m^{6}A in active gene regulation may manifest itself as altered methylation profiles in response to changing cellular conditions or in a tissue-specific manner.

We put our hypothesis to the test by comparing the methylation profiles of untreated HepG2 cells to those of cells exposed to ultraviolet radiation, heat shock, hepatocyte growth factor (HGF; also known as scatter factor (SF)), and interferon-γ (Supplementary Table 2). Remarkably, all samples exhibited a marked similarity of

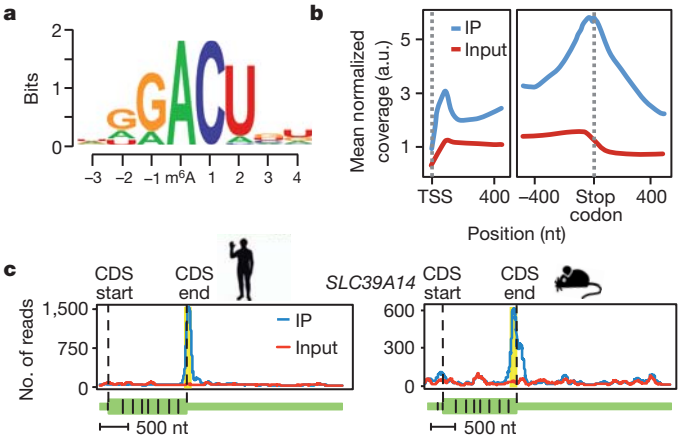


Figure 3 | m^{6}A methylome conservation between human and mouse. **a**, Sequence logo representing the deduced consensus motif following clustering of all enriched motifs. **b**, Metagene profiles depicting sequence coverage in windows surrounding the TSS (left) and stop codon (right). Colour codes are the same as in panel Fig. 2a. The mouse stop codon segment is even more enriched in m^{6}A peaks than its human counterpart, with 39% of all peaks located in this segment, corresponding to >3.5-fold enrichment over the distribution expected by chance ($P < 2.2 \times 10^{-308}$, χ^2 test). **c**, Orthologous genes with conserved m^{6}A peaks. Gene architecture is shown beneath. a.u., arbitrary units.

m^{6}A profiles, with 70–95% peak positions typically shared between conditions (Supplementary Fig. 11).

Nevertheless, we were able to detect a subset of treatment-dependent, dynamically altered peaks (Fig. 4 and Supplementary Table 8). These peaks did not correlate with absence of a proximal consensus motif. We note that our stringent approach probably underestimates the amount of differential m^{6}A peaks, all the more as it is insensitive to the proportion of methylated versions of a specific transcript.

We further examined tissue specificity and compared normal human brain to HepG2 cells. The former recapitulated all the global features of the m^{6}A methylome initially identified in HepG2 (including consensus motif and peak distribution along gene architecture) to demonstrate that the characteristics of the methylome are not an aberration of cancer tissues but rather a normal phenomenon (Supplementary Note 2).

m^{6}A affects RNA splicing

The function of m^{6}A in RNA metabolism was next assessed by silencing of *METTL3* in HepG2 cells, as its depletion was already shown to reduce the amount of m^{6}A in the transcriptome[4]. *METTL3* knockdown resulted in apoptosis (Supplementary Fig. 12), as expected[3]. Reads obtained from massively parallel sequencing of control (mock) and knockdown RNA were aligned to the genome and differential gene expression profiles were generated (Methods; Supplementary Table 4). Of 1,977 differentially expressed genes, 1,218 contain mapped m^{6}A peaks. Downregulated genes were significantly enriched with genes shown to have methylated introns ($P = 1.3 \times 10^{-7}$, hypergeometric test; Supplementary Fig. 13).

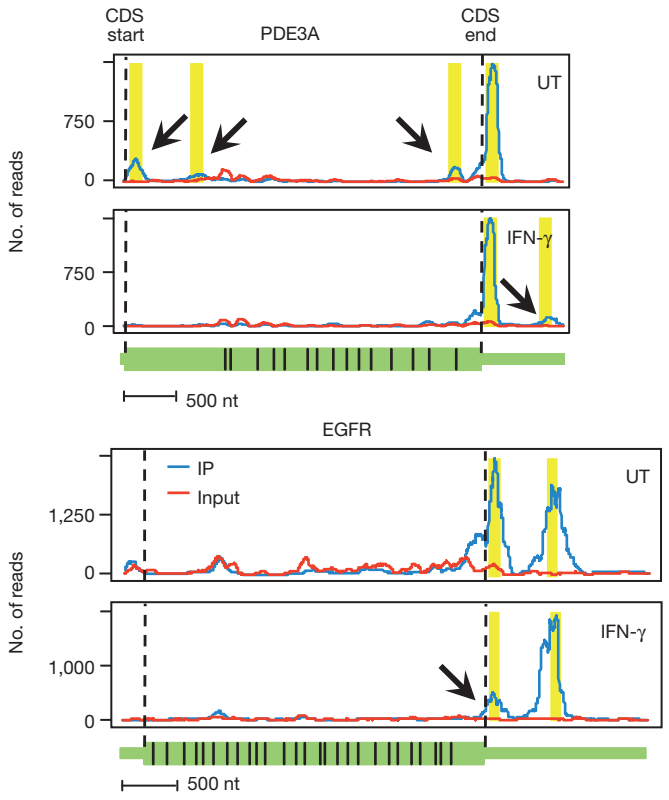


Figure 4 | Transcripts differing in methylation patterns under varying growing conditions. Arrows point at the differential position. The stability of the surrounding peaks underscores the authenticity of the changed peaks. Interferon-γ (IFN-γ) upregulates epidermal growth factor receptor (EGFR) RNA and protein levels[47,48], suggesting an association with absence of m^{6}A from the stop codon segment. PDE3A, phosphodiesterase 3A, cGMP inhibited; UT, untreated.

We reasoned that focusing on genes that were not differentially expressed (fold change < 2) but whose constituent isoforms were (fold change > 2), would help isolate a possible correlation between m^{6}A and isoform switching (between up- and downregulated isoforms). Indeed, methylated genes were overrepresented in this set (459/543, $P = 7.9 \times 10^{-12}$, hypergeometric test). Moreover, differentially spliced exons and introns were themselves significantly enriched with m^{6}A peaks: 99/474 exons ($P = 4 \times 10^{-15}$, hypergeometric test) and 811/2,672 introns ($P = 1 \times 10^{-105}$, hypergeometric test), further supporting a role for m^{6}A in splicing.

The salient findings revealed by *METTL3* knockdown prompted us to examine all m^{6}A peaks in the HepG2 transcriptome through the alternative splicing prism. Assigning peaks to either single- or multi-isoform coding genes revealed that the former are relatively undermethylated (6,489 and 11,946 peaks in 3,698 and 5,870 coding genes, respectively; $P = 5.6 \times 10^{-7}$, hypergeometric test). Accordingly, the average number of m^{6}A peaks per coding gene (1.93 peaks per gene) was higher in multi-isoform genes (2.04 peaks per gene) than in single-isoform ones (1.75 peaks per gene). Refining this observation further, assignment of all m^{6}A peaks to either constitutive or alternative spliced exons revealed that although 23.3% of coding exons are constitutive, they are significantly under-represented among all methylated sequences (10.9%, $P < 1 \times 10^{-305}$, hypergeometric test; Supplementary Fig. 14).

Interestingly, GO analysis of differentially expressed genes indicated a noteworthy enrichment of the *p53* signalling pathway (23 genes, corrected $P = 6.0 \times 10^{-5}$): 22/23 genes had differentially expressed splice variants, of which 18 were methylated. Moreover, 15 other members of the signalling pathway, which did not show significant differential expression at the gene level, exhibited significant differential expression at the isoform level (Supplementary Fig. 15a). For example, isoforms of *MDMX* (also known as *MDM4*), needed for p53 inactivation[30], were downregulated (Supplementary Fig. 15b). Similar pro-apoptotic effects were observed in other key genes belonging to this pathway (for example, *MDM2*, *FAS* and *BAX*). Modulation of *p53* signalling through splicing may be relevant to induction of apoptosis by silencing of *METTL3*.

Knockout of *IME4*, the yeast orthologue of *METTL3*, demonstrated its important role in regulation of the developmental switch from vegetative cells into gametogenesis, but failed to provide mechanistic insight (Supplementary Fig. 16).

Binding of RNA-interacting proteins

Methylation of RNA may also affect binding of interacting proteins, similar to recognition of 5-methylcytosine in DNA by specific binding proteins that mediate its repressive effects[31]. We observed an overlap between known RNA sequence elements and newly identified m^{6}A peaks: several internal ribosome entry sites (IRES)[32], the localization 'zip code' of β-actin[33] and the destabilization element of *c-Myc* (also known as *MYC*)[34], all harbour methylation peaks (Supplementary Fig. 2b, c and Supplementary Fig. 17).

An RNA affinity chromatography approach, using methylated and control versions of an RNA bait followed by mass spectrometry, was used to identify novel m^{6}A-binding proteins (Supplementary Fig. 18a).

Our analysis identified three RNA-binding proteins that may mediate novel connections between m^{6}A and cellular processes (Supplementary Fig. 18b and Supplementary Table 5). Two YTH (YT521-B homology) family proteins, *YTHDF2* and *YTHDF3*, bound exclusively to the methylated bait (Supplementary Fig. 18b, d). These proteins contain a recently characterized RNA-binding domain (YTH) that overlaps the methylated motif in our bait[35] (Supplementary Fig. 18c). Interestingly, the only two characterized YTH family members, the human *YT521-B* (also known as *YTHDC1*) and the fission yeast *Mmi1*, were implicated in alternative splice site selection and in specifying transcripts for nuclear degradation, respectively[36,37].

Another protein that was significantly associated with the m^{6}A bait was *ELAVL1* (also known as *HUR*)[38] (Supplementary Fig. 18b, d).

Discussion

Lack of knowledge regarding the distribution of m^{6}A in the transcriptome was one of the major factors that limited understanding of its role in RNA metabolism. m^{6}A-seq provides a valuable tool to bridge this gap and allows global mapping of the methylome to uncover several of its fundamental properties. The human and mouse RNA methylomes presented here reveal that RNA is non-randomly punctuated by m^{6}A throughout the body of most transcripts. We observe that m^{6}A sites are correlated with two distinct landmarks with telling biological importance: around stop codons and within long internal exons. The fact that this architectural pattern is also the most evolutionary conserved feature of the RNA methylome indicates an elementary role for this modification.

The unique distribution of m^{6}A provides hints as to its functions. It is unnecessary to assume one unifying principle behind this modification—m^{6}A can be involved in several cellular functions, similar to inosine, the best-studied global RNA modification[39]. The discovery of proteins that preferentially bind to methylated RNA raises the possibility that they are involved in mediating its biological effects.

Methylation of long internal exons is suggestive of involvement in splicing. Perhaps it reflects an auxiliary measure to alleviate the constraints imposed by larger exons on the exon-definition machinery[40,41]. The correlation of methylation with multi-isoform genes and differentially spliced exons and introns indicates a role for m^{6}A in splicing control. In line with this hypothesis, *METTL3* and *FTO* co-localize with splicing proteins in nuclear speckles[4,16], and the application of methylation inhibitors to cells resulted in nuclear accumulation of unspliced transcripts[42–44]. Integration of the methylation layer into 'splicing codes', such as that previously described[45], may improve the predictive power of alternative splicing events.

Methylation in the vicinity of stop codons hints at the direction of translational control. It is tempting to speculate that the presence of m^{6}A around this landmark affects translation efficiency, either directly or through the recruitment of specific factors. Indeed, the link between m^{6}A and translational efficiency was previously demonstrated *in vitro*[46].

The dynamic regulation of nucleic-acid post-transcriptional modifications, analogous to post-translational protein modifications, is just beginning to be elucidated[28]. Here we provide initial evidence for a change in m^{6}A positions and frequencies in response to various perturbations. Given the previously established non-stoichiometric nature of this base modification[9] and existence of the *FTO* demethylase[16], it seems reasonable to speculate a role for it as part of a global cellular response to stimuli.

The m^{6}A methylome opens new avenues for correlating the methylation layer with other processing levels. In many ways, this approach is a forerunner, providing a reference and paving the way for the uncovering of other RNA modifications, which together constitute a new realm of biological regulation, recently termed RNA epigenetics[1].

METHODS SUMMARY

RNA was randomly fragmented and subjected to immunoprecipitation using an anti-m^{6}A antibody. Immunoprecipitated and input control samples were sequenced using Illumina GAIIx. Using both in-house and MACS peak-calling algorithms, regions enriched in immunoprecipitated relative to input samples were identified as m^{6}A peaks comprising the methylome. Motifs enriched around m^{6}A peaks were identified. Human and mouse methylomes were compared by analysing methylation profiles of orthologous genes. *METTL3* knockdown was achieved by short interfering RNA (siRNA) transfection. Control and *METTL3* knockdown samples were sequenced using Illumina GAIIx and subjected to differential expression analyses. RNA affinity chromatography experiments were carried out using two biotinylated RNA baits differing in a single m^{6}A modification and bound proteins were analysed by mass spectrometry (LC-MS/MS).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 April 2011; accepted 11 April 2012.
Published online 29 April 2012.

1. He, C. Grand challenge commentary: RNA epigenetics? *Nature Chem. Biol.* **6,** 863–865 (2010).
2. Cantara, W. A. *et al.* The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39,** D195–D201 (2011).
3. Bokar, J. A. in *Fine-Tuning of RNA Functions by Modification and Editing* Vol. 12 (ed. Grosjean, H.) 141–177 (Springer, 2005).
4. Bokar, J. A., Shambaugh, M. E., Polayes, D., Matera, A. G. & Rottman, F. M. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N^6-adenosine)-methyltransferase. *RNA* **3,** 1233–1247 (1997).
5. Zhong, S. *et al.* MTA is an *Arabidopsis* messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *Plant Cell* **20,** 1278–1288 (2008).
6. Clancy, M. J., Shambaugh, M. E., Timpte, C. S. & Bokar, J. A. Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of *N*^{6}-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res.* **30,** 4509–4518 (2002).
7. Hongay, C. F. & Orr-Weaver, T. L. *Drosophila* Inducer of MEiosis 4 (IME4) is required for Notch signaling during oogenesis. *Proc. Natl Acad. Sci. USA* **108,** 14855–14860 (2011).
8. Desrosiers, R., Friderici, K. & Rottman, F. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proc. Natl Acad. Sci. USA* **71,** 3971–3975 (1974).
9. Horowitz, S., Horowitz, A., Nilsen, T. W., Munns, T. W. & Rottman, F. M. Mapping of N6-methyladenosine residues in bovine prolactin mRNA. *Proc. Natl Acad. Sci. USA* **81,** 5667–5671 (1984).
10. Kane, S. E. & Beemon, K. Precise localization of m6A in Rous sarcoma virus RNA reveals clustering of methylation sites: implications for RNA processing. *Mol. Cell. Biol.* **5,** 2298–2306 (1985).
11. Harper, J. E., Miceli, S. M., Roberts, R. J. & Manley, J. L. Sequence specificity of the human mRNA N6-adenosine methylase *in vitro*. *Nucleic Acids Res.* **18,** 5735–5741 (1990).
12. Wei, C. M. & Moss, B. Nucleotide sequences at the N^6-methyladenosine sites of HeLa cell messenger ribonucleic acid. *Biochemistry* **16,** 1672–1676 (1977).
13. Dai, Q. *et al.* Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and *N*^{6}-methyladenosine. *Nucleic Acids Res.* **35,** 6322–6329 (2007).
14. Levanon, E. Y. *et al.* Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nature Biotechnol.* **22,** 1001–1005 (2004).
15. Perry, R. P. & Scherrer, K. The methylated constituents of globin mRNA. *FEBS Lett.* **57,** 73–78 (1975).
16. Jia, G. *et al.* *N*^{6}-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chem. Biol.* **7,** 885–887 (2011).
17. Bringmann, P. & Luhrmann, R. Antibodies specific for *N*^{6}-methyladenosine react with intact snRNPs U2 and U4/U6. *FEBS Lett.* **213,** 309–315 (1987).
18. Dante, R. & Niveleau, A. Inhibition of *in vitro* translation by antibodies directed against *N*^{6}-methyladenosine. *FEBS Lett.* **130,** 153–157 (1981).
19. Munns, T. W., Liszewski, M. K., Oberst, R. J. & Sims, H. F. Antibody nucleic acid complexes. Immunospecific retention of N6-methyladenosine-containing transfer ribonucleic acid. *Biochemistry* **17,** 2573–2578 (1978).
20. Munns, T. W., Liszewski, M. K. & Sims, H. F. Characterization of antibodies specific for N^6-methyladenosine and for 7-methylguanosine. *Biochemistry* **16,** 2163–2168 (1977).
21. Munns, T. W., Oberst, R. J., Sims, H. F. & Liszewski, M. K. Antibody-nucleic acid complexes. Immunospecific recognition of 7-methylguanine- and *N*^{6}-methyladenine-containing 5′-terminal oligonucleotides of mRNA. *J. Biol. Chem.* **254,** 4327–4330 (1979).
22. Munns, T. W., Sims, H. F. & Liszewski, M. K. Immunospecific retention of oligonucleotides possessing *N*^{6}-methyladenosine and 7-methylguanosine. *J. Biol. Chem.* **252,** 3102–3104 (1977).
23. Czerwoniec, A. *et al.* MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res.* **37,** D118–D121 (2009).
24. Perry, R. P., Kelley, D. E., Friderici, K. & Rottman, F. The methylated constituents of L cell messenger RNA: evidence for an unusual cluster at the 5′ terminus. *Cell* **4,** 387–394 (1975).
25. Keith, J. M., Ensinger, M. J. & Mose, B. HeLa cell RNA (2′-*O*-methyladenosine-*N*^{6}-)-methyltransferase specific for the capped 5′-end of messenger RNA. *J. Biol. Chem.* **253,** 5033–5039 (1978).
26. Wei, C., Gershowitz, A. & Moss, B. *N*^{6}, *O*$^{2'}$-dimethyladenosine a novel methylated ribonucleoside next to the 5′ terminal of animal cell and virus mRNAs. *Nature* **257,** 251–253 (1975).
27. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genet.* **39,** 61–69 (2007).
28. Chan, C. T. *et al.* A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet.* **6,** e1001247 (2010).
29. Schaefer, M. *et al.* RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.* **24,** 1590–1595 (2010).
30. Lenos, K. & Jochemsen, A. G. Functions of MDMX in the modulation of the p53-response. *J. Biomed. Biotechnol.* **2011,** 876173 (2011).

31. Klose, R. J. & Bird, A. P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31,** 89–97 (2006).
32. Mokrejs, M. *et al.* IRESite—a tool for the examination of viral and cellular internal ribosome entry sites. *Nucleic Acids Res.* **38,** D131–D136 (2009).
33. Kislauskis, E. H., Zhu, X. & Singer, R. H. Sequences responsible for intracellular localization of β-actin messenger RNA also affect cell phenotype. *J. Cell Biol.* **127,** 441–451 (1994).
34. Bernstein, P. L., Herrick, D. J., Prokipcak, R. D. & Ross, J. Control of c-myc mRNA half-life *in vitro* by a protein capable of binding to a coding region stability determinant. *Genes Dev.* **6,** 642–654 (1992).
35. Zhang, Z. *et al.* The YTH domain is a novel RNA binding domain. *J. Biol. Chem.* **285,** 14701–14710 (2010).
36. Harigaya, Y. *et al.* Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature* **442,** 45–50 (2006).
37. Rafalska, I. *et al.* The intranuclear localization and function of YT521-B is regulated by tyrosine phosphorylation. *Hum. Mol. Genet.* **13,** 1535–1549 (2004).
38. Brennan, C. M. & Steitz, J. A. HuR and mRNA stability. *Cell. Mol. Life Sci.* **58,** 266–277 (2001).
39. Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79,** 321–349 (2010).
40. Keren, H., Lev-Maor, G. & Ast, G. Alternative splicing and evolution: diversification, exon definition and function. *Nature Rev. Genet.* **11,** 345–355 (2010).
41. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *Proc. Natl Acad. Sci. USA* **93,** 15081–15085 (1996).
42. Camper, S. A., Albers, R. J., Coward, J. K. & Rottman, F. M. Effect of undermethylation on mRNA cytoplasmic appearance and half-life. *Mol. Cell. Biol.* **4,** 538–543 (1984).
43. Carroll, S. M., Narayan, P. & Rottman, F. M. N6-methyladenosine residues in an intron-specific region of prolactin pre-mRNA. *Mol. Cell. Biol.* **10,** 4456–4465 (1990).
44. Stoltzfus, C. M. & Dane, R. W. Accumulation of spliced avian retrovirus mRNA is inhibited in *S*-adenosylmethionine-depleted chicken embryo fibroblasts. *J. Virol.* **42,** 918–931 (1982).
45. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
46. Tuck, M. T., Wiehl, P. E. & Pan, T. Inhibition of 6-methyladenine formation decreases the translation efficiency of dihydrofolate reductase transcripts. *Int. J. Biochem. Cell Biol.* **31,** 837–851 (1999).
47. Hamburger, A. W. & Pinnamaneni, G. Interferon-induced enhancement of transforming growth factor-α expression in a human breast cancer cell line. *Proc. Soc. Exp. Biol. Med.* **202,** 64–68 (1993).
48. Mujoo, K., Donato, N. J., Lapushin, R., Rosenblum, M. G. & Murray, J. L. Tumor necrosis factor α and γ-interferon enhancement of anti-epidermal growth factor receptor monoclonal antibody binding to human melanoma cells. *J. Immunother. Emphasis Tumor Immunol.* **13,** 166–174 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Cedar (The Hebrew University, Jerusalem) for his comments. We thank the Kahn Family Foundation for their support. This work was supported in part by grants from the Flight Attendant Medical Research Institute (FAMRI), Bio-Med Morasha ISF (grant no. 1942/08), and The Israel Ministry for Science and Technology (Scientific Infrastructure Program). R.S. was supported by the ERC-StG program (grant 260432). G.R. holds the Djerassi Chair in Oncology at the Sackler Faculty of Medicine, Tel Aviv University. This work was performed in partial fulfilment of the requirements for a PhD degree to D.D., Sackler Faculty of Medicine, Tel Aviv University.

Author Contributions D.D., S.M.-M. and G.R. conceived and designed the experiments; D.D., S.M.-M., L.U., K.C., S.O. and J.J.-H. performed the experiments; S.S., M.S.-D. and R.S. performed the bioinformatic analysis; D.D., S.M.-M., M.S.-D., N.A., M.K., S.S., R.S. and G.R. analysed and interpreted results, and wrote the paper.

Author Information Data have been deposited in NCBI's Gene Expression Omnibus (GEO) and are accessible through GEO series accession number GSE37005 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37005). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to G.R. (Gidi.Rechavi@sheba.health.gov.il).

METHODS

Tissues, cell lines and strains. Human hepatocellular carcinoma cell line, HepG2, was maintained in DMEM (Gibco, Invitrogen) containing 4.5 g l^{-1} glucose and L-glutamine supplemented with 10% FBS and penicillin/streptomycin. Where indicated, HepG2 cells were incubated with IFN-γ (200 ng ml^{-1}) or HGF (10 ng ml^{-1}) overnight. Stress effects were tested in HepG2 cells by either 30 min incubation at 43 °C (heat shock) or ultraviolet irradiation of 0.04 J cm^{-2} at 254 nm. Cells were allowed to recover for 4 h in normal growing conditions before harvesting using trypsin.

Yeast. Wild-type and *ime4Δ* mutant (created by conventional knockout protocols) SK1 *S. cerevisiae* cells were grown vegetatively in YPD medium (yeast extract, peptone, dextrose), and transferred to sporulation medium (1% potassium acetate) for induction of sporulation for 4 h. RNA was purified from the cells using TRIzol (Invitrogen) and glass beads, and hybridized to Affymetrix microarrays. Experiments were conducted in biological replicates.

Microarrays. Biotinylated cDNA was prepared according to standard Affymetrix protocol from 600 ng total RNA. After fragmentation, 10 μg of cRNA were hybridized for 16 h at 45 °C on yeast 2.0 Genome Array. GeneChips were washed and stained in the Affymetrix Fluidics Station 400. Feature intensity values from scanned arrays were normalized and reduced to expression summaries using the RMA algorithm. Control probes and probe sets with an interquartile range below 0.1 were removed.

RNA preparation. Total human brain RNA was purchased from Biochain. Total RNA from HepG2 cell pellets was extracted using PerfectPure RNA cultured cell kit (5 Prime). Total RNA of C57/BL6 mouse livers was extracted using PerfectPure RNA tissue kit (5 Prime). To avoid DNA contaminations all samples were treated with DNase (5 Prime).

Enrichment of polyadenylated RNA (polyA$^+$ RNA) from total RNA was performed using one round of GenElute mRNA miniprep kit (Sigma-Aldrich).

RNA samples were chemically fragmented into ~100-nucleotide-long fragments by 5 min incubation at 94 °C in fragmentation buffer (10 mM $ZnCl_2$, 10 mM Tris-HCl, pH 7). The fragmentation reaction was stopped with 0.05 M EDTA, followed by standard ethanol precipitation. Samples were resuspended in H_2O at ~1 μg μl^{-1} concentration and subjected to m^{6}A-seq.

m^{6}A-seq. Fragmented RNA (400 μg mRNA or 2.5 mg total RNA) was incubated for 2 h at 4 °C with 5 μg of affinity purified anti-m^{6}A polyclonal antibody (Synaptic Systems) in IPP buffer (150 mM NaCl, 0.1% NP-40, 10 mM Tris-HCl, pH 7.4). The mixture was then immunoprecipitated by incubation with protein-A beads (Repligen) at 4 °C for an additional 2 h. After extensive washing, bound RNA was eluted from the beads with 0.5 mg ml^{-1} N^6-methyladenosine (Sigma-Aldrich) in IPP buffer, and ethanol precipitated. RNA was resuspended in H_2O and used for library generation with mRNA sequencing kit (Illumina). Sequencing was carried out on Illumina Genome Analyzer (GAIIx) according to the manufacturer's instructions, using a 10 pM template per sample for cluster generation, and genomic sequencing kit V2 (Illumina).

Gene silencing by siRNA. Scrambled control or siRNA directed against *METTL3* (Invitrogen, 5′-AUAUCACAACAGAUCCACUGAGGUG-3′) were transfected into HepG2 cells using oligofectamine reagent (Invitrogen). Cells were re-transfected after 48 and 96 h, and harvested on days 2, 5 and 7, to allow more efficient METTL3 silencing. RNA purification and massively parallel sequencing were performed as described above. The experiment was conducted in two biological replicates, generating ~55 and ~70 million reads in knockdown and control respectively.

TUNEL assay. METTL3 knockdown and mock control cells were assayed for apoptosis on day 7 using DeadEnd Fluorometric TUNEL System (Promega). Nuclei were stained with DAPI (Sigma-Aldrich). Samples were immediately visualized using LSM 510 confocal microscope (Zeiss).

RNA affinity chromatography, mass spectrometry and western blot analysis. Two biotin-labelled RNA oligonucleotide baits spanning 42 nucleotides centred on a previously characterized m^{6}A site in the RSV genome were synthesized, only one of which was N^6-methylated: 5′-biotin-AUGGGCCGUUCAUCUGCUAAAAGG-m^{6}A- CUGCUUUUGGGGCUUGU-3′ and 5′-biotin-AUGGGCCGUUCAUCUGCUAAAAGGACUGCUUUUGGGGCUUGU-3′. HepG2 cells were harvested at 70–80% confluence, washed with PBS and lysed in lysis buffer (10 mM NaCl, 2 mM EDTA,0.5% Triton X-100, 0.5 mM DTT, 10 mM Tris-HCl, pH 7.5) containing complete protease inhibitor cocktail (Roche) and phosphatase inhibitor cocktail 2 (Sigma-Aldrich). Lysates were separated from insoluble cell debris by centrifugation (10,000*g* for 15 min at 4 °C) and pre-cleared for 1 h at 4 °C by incubation with streptavidin-conjugated agarose beads (Sigma-Aldrich) in binding buffer (150 mM KCl, 1.5 mM $MgCl_2$, 0.05% (v/v) NP-40, 0.5 mM DTT, 10 mM Tris-HCl pH 7.5). Biotinylated RNA baits (2 μg) were incubated with pre-cleared cell lysates supplemented with 0.4 units μl^{-1} RNasin (Promega) for 30 min at room temperature followed by 2 h incubation at 4 °C. The mixture was then added to streptavidin-conjugated agarose beads pre-blocked with BSA (1%) and tRNA (50 μg ml^{-1}) for 2 h at 4 °C. RNA–protein complexes were pulled-down and washed extensively. Samples were separated on 10% (w/v) polyacrylamide Bis-Tris gels (Invitrogen) and stained with Imperial protein stain (Thermo scientific). Proteins in gel slices were digested with trypsin and identified using LC-MS/MS by the Weizmann Institute's Mass Spectrometry Unit. Three independent biological replicates were performed. For western blot analyses samples were separated on 10% (w/v) polyacrylamide Bis-Tris gels (Invitrogen) and transferred onto nitrocellulose membrane using iBlot gel transfer system (Invitrogen) set to P3 for 8 min with iBlot gel transfer stacks (Invitrogen). Membranes were blocked in 5% BSA, 0.05% Tween-20 in PBS for 1 h, and then incubated overnight at 4 °C with anti-YTHDF2 or anti-ELAVL1 polyclonal antibody (Abcam) diluted 1:500 in 5% milk; anti-METTL3 (Abnova) diluted 1:2,000; anti-actin (Epitomics) diluted 1:5,000. Proteins were visualized using the SuperSignal West Pico Luminol/Enhancer solution (Thermo scientific).

Alignment of reads. Transcript-based analysis. Reference transcriptomes were prepared based on the University of California, Santa Cruz (UCSC) KnownCanonical tables in human (hg18) and mouse (mm9) consisting of a single representative transcript for each gene[49], and comprising 19,965 coding and 5,950 non-coding genes for human, and 20,437 and 6,739, respectively, for mouse. Reads were first aligned against the relevant reference using Novoalign 2.07 (Novocraft Technologies SdnBhd; http://www.novocraft.com), discarding all reads that were mapped to multiple genomic regions. The remaining reads were realigned against the relevant transcriptome and reads that aligned less well to the transcriptome than to the genome were also discarded. A strategy of iteratively trimming two nucleotides from the end of the read and then realigning it was used when a read did not align at its original length (-s 2 parameter). Trimming of reads to a length shorter than 25 nucleotides (-l 25 parameter) was not allowed. Additional parameters used for the indexing step were '-k 14 -s 2', and for the alignment was '-t 75'.

Genome-based analysis. Reads were mapped to the human genome using Bowtie[50] with '-best -strata' parameters. Non-unique reads mapping to more than ten locations were discarded from downstream analysis.

Calculation of coverage. Reads were artificially extended to a length of 100 nucleotides in the 5′-to-3′ direction, to account for the average length of RNA fragments, thus ensuring coverage of methylation sites residing outside the sequenced 36-nucleotide stretch. The extended reads were used to generate gene coverage maps summing the number of reads overlapping each nucleotide in every gene. This approach relies on the narrow fragment size range (90–150 nucleotides, measured by Agilent Bioanalyzer). Analyses were limited to human genes with at least 40 reads per 1,000 nucleotides, to avoid misinterpretations due to insufficient coverage. Expression levels were evaluated by calculating the number of reads in each gene, and dividing the range into five bins.

Detection of m^{6}A sites. Search for enriched peaks in the m^{6}A immunoprecipitation sample compared to the input control was performed by scanning each gene using sliding windows of 100 nucleotides with 50 nucleotides overlap. The mean coverage for each window was calculated for the immunoprecipitated and control samples (MeanWinIP and MeanWinControl, respectively). Gene median coverages for immunoprecipitation and input control were determined (MedianGeneIP and MedianGeneControl, respectively) to robustly estimate background levels. Every window was assigned the following unit-less metric, representing the enrichment fold change of immunoprecipitated over input control samples after normalization by background.

$$\text{winscore} = \log_2\left(\frac{\text{MeanWinIP}/\text{MedianGeneIP}}{\text{MeanWinControl}/\text{MedianGeneControl}}\right)$$

A robust estimation signal was ensured by setting a limit of mean window coverage >20 for the immunoprecipitated sample. The FDR of this method was estimated by comparing the number of enriched windows exceeding a set threshold using this approach to the number that is obtained if the experiments were computationally reversed with input being treated as immunoprecipitate and vice versa. For a threshold of 2, corresponding to a fourfold increase with respect to control, the FDR was <7%; thus, the winscore threshold was set to 2. All overlapping enriched windows were merged and positions with maximal immunoprecipitate-positive and immunoprecipitate-negative coverage were identified. If the distance between these two maxima was <100 nucleotides and the ratio between the amplitude of these two maxima was less than twofold, a putative m^{6}A site was defined in the centre of these two maxima, and a score was allocated to this peak, peakscore, corresponding to the maximal winscore from among all the overlapping windows. When these conditions were not met, windows were classified as non-resolved peaks and excluded from the analysis, to avoid PCR/sequencing artefacts. The negative control peaks were detected as described above: All windows with winscore <0 were selected. Control windows were

selected from the same genes as the detected m^{6}A peaks (limited to a mean window coverage >20). All overlapping control windows matching these criteria were merged and a negative control peak was defined in their centre.

For genome-based analysis, the peak-caller MACS[51] was used for peak detection; the 'effective genome size' parameter was adjusted to the calculated transcriptome size (1.35×10^8). Peaks were considered if their MACS-assigned fold change was >4 and individual FDR value <5%.

Assignment of m^{6}A peaks into exons representing different splicing events. m^{6}A peak summits were intersected with a data set of genomic exon coordinates belonging to ten splicing categories. A list of splicing event exons of all coding Ensembl genes (version 63 hg19) was downloaded using the BioMart tool[52]. After lifting over peak summits to hg19 coordinates, they were intersected with the downloaded splicing event list.

Identification and clustering of enriched motifs. In-house method for detecting motifs using all identified peaks. Motifs enriched within m^{6}A peaks compared with control peaks were identified by counting the occurrence of 4–6-nucleotide *k*-mers in the immunoprecipitate and its corresponding control group. The total number of *k*-mers of each length within every group was counted and the ratio between their prevalence was used to calculate the fold change between the two groups. Exact Fisher test was used to evaluate the differences in the prevalence of each *k*-mer between the groups. Analysis was limited to motifs enriched more than twofold and with an associated Bonferonni-corrected *P* value < 0.05. Motifs were clustered together, using the previously described approach[53]. To correct for the underlying base composition we repeated these analyses using a second set of control sequences based on randomly permuting the sequences of the m^{6}A sites, validating the significance of the found motif.

MEME search. MACS-identified peaks with FDR $\leq$ 5% were sorted according to their fold change. The top 1,000 peaks falling within known genes were chosen for *de novo* motif analysis. 101-nucleotide-long sequences derived from the sense strand and centred around the peak summit were used as input for MEME[54].

Conservation analysis. Human–mouse orthologous genes were identified based on the mmBlastTab table, which is linked to the UCSC KnownCanonical table. Orthologous sequences were then aligned using the Needleman–Wunsch global alignment algorithm implemented in 'needle'. Mouse and human m^{6}A peaks were identified between their start and end coordinates (see above) and projected onto the alignment to enable the identification of conserved m^{6}A peaks. The distance of the intervals in human was limited to be >50% the length in mouse, to ensure minimal sequence identity; sites not matching this criterion were discarded. Analyses were limited to human genes with at least 40 reads per 1,000 nucleotides, to avoid misinterpretations due to insufficient coverage.

m^{6}A stability across experiments. To explore the stability of m^{6}A peaks across different experiments, the presence of each peak identified in a particular (reference) experiment was confirmed in the other (target) experiments. Conserved peaks were defined as peaks existing in both experiments between matching coordinates. Only peaks in genes with a mean expression exceeding 40 reads per 1,000 nucleotides in the target experiment were considered, to guarantee sufficient coverage. Borderline cases passing the threshold in one experiment but not in the other were minimized by limiting the peaks in the reference experiment to a peakscore of >3. For the analysis of differential peaks between two experiments, we demanded that the peakscore be <0 in the target experiment lacking the site, to ensure the presence of a peak in only one condition.

Secondary structure. RNAfold[55] was used to assess secondary structures within 100-nucleotide sequences centred around the detected m^{6}A and control peaks with default parameters to calculate the minimum free energy (MFE) of the folded sequence. Each sequence was also shuffled 50 times and MFE scores were calculated for those sequences as well. Sequences were then assigned a *Z* score, indicative of the extent to which a sequence is more stably folded compared to the shuffled controls.

Multi-layer differential expression analysis. For all differential expressed layers, features were considered as differentially expressed only when fold change ≥ 2 and FDR $\leq$ 5% using Benjamini–Hochberg multiple testing adjustment.

Differential expressed genes. RNA-seq reads were mapped to the human genome (build hg18) using TopHat 1.2.0 software[56]. The number of reads mapped to each of the Ensembl genes (release 54) was counted using the HTSeq python package[57], with the 'union' overlap resolution mode, and -stranded = no. The R package DESeq v.1.5.24[58] within the Bioconductor framework was used for differential expression analysis.

Differentially expressed exons. The analysis was done using the R package DEXSeq[59] within the Bioconductor framework. The package provides a method to systematically detect differential exon usage, that is, whether the proportion of read from a given exon among all the reads that fall onto a gene is significantly changed. The analysis was done in the following way. First, non-overlapping exonic regions were defined using the 'dexseq_prepare_annotation.py' script provided as part of the DEXseq package. Next the number of reads falling in each of the defined exonic regions was counted using the DEXseq script 'dexseq_count.py' with parameters -a = 0 to include multi-reads mapped to different locations of the genome, and -stranded = no.

Differentially expressed introns. Non-overlapping intronic regions that also do not overlap any exon on either strand were defined using an in-house script. The number of reads falling in each of the defined regions was counted using a modified version of the 'dexseq_count.py' DEXseq script, and differential expression analysis was done using the DESeq Bioconductor package.

Differentially expressed transcripts. An Ensembl .gtf file of all human genes (hg18 release 54) was re-processed using Cuffcompare v.1.0.3 to add the missing tss_id and p_id attributes according to the user guide. The resulting .gtf annotation file created by Cuffcompare was used as input to Cuffdiff v.1.0.3 tool together with the fragment alignment files. Both Cuffcompare and Cuffdiff are part of the Cufflinks package[60].

Intersection between data sets. Intersection operations between genomic locations of differentially expressed features and the identified peaks were done using BEDTools[61] and PeakAnalyzer software[62]. Hypergeometric test was used for calculating enrichment *P* values, unless stated otherwise.

Statistical analysis and graphics. All statistical analyses (unless stated otherwise) were performed using the R package for Statistical Computing. Most of the presented figures were produced using the ggplot2 package[63]. Sequence logos were prepared using the SeqLogo package[64].

49. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22,** 1036–1046 (2006).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
51. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
52. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011,** bar030 (2011).
53. Llorian, M. *et al.* Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Struct. Mol. Biol.* **17,** 1114–1123 (2010).
54. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. ISMB* **2,** 28–36 (1994).
55. Zuker, M. & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9,** 133–148 (1981).
56. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
57. Anders, S. HTSeq: analysing high-throughput sequencing data with Python http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html (2010).
58. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
59. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-Seq data. Available from *Nature Precedings* http://hdl.handle.net/10101/npre.2012.6837.1 (2012).
60. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
61. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
62. Salmon-Divon, M., Dvinge, H., Tammoja, K. & Bertone, P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. *BMC Bioinformatics* **11,** 415 (2010).
63. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
64. Bembom, O. *seqLogo: Sequence Logos for DNA Sequence Alignments* (Division of Biostatistics, University of California, Berkeley, 2011).

ARTICLE

doi:10.1038/nature11010

Molecular mechanism of ATP binding and ion channel activation in P2X receptors

Motoyuki Hattori[1] & Eric Gouaux[1,2]

P2X receptors are trimeric ATP-activated ion channels permeable to Na^+, K^+ and Ca^{2+}. The seven P2X receptor subtypes are implicated in physiological processes that include modulation of synaptic transmission, contraction of smooth muscle, secretion of chemical transmitters and regulation of immune responses. Despite the importance of P2X receptors in cellular physiology, the three-dimensional composition of the ATP-binding site, the structural mechanism of ATP-dependent ion channel gating and the architecture of the open ion channel pore are unknown. Here we report the crystal structure of the zebrafish $P2X_4$ receptor in complex with ATP and a new structure of the apo receptor. The agonist-bound structure reveals a previously unseen ATP-binding motif and an open ion channel pore. ATP binding induces cleft closure of the nucleotide-binding pocket, flexing of the lower body β-sheet and a radial expansion of the extracellular vestibule. The structural widening of the extracellular vestibule is directly coupled to the opening of the ion channel pore by way of an iris-like expansion of the transmembrane helices. The structural delineation of the ATP-binding site and the ion channel pore, together with the conformational changes associated with ion channel gating, will stimulate development of new pharmacological agents.

ATP, best known for its roles in energy metabolism, intracellular signalling, biosynthetic reactions and active transport, is a crucial extracellular signalling molecule[1] that binds to two different classes of ATP receptors: ionotropic P2X receptors[2] and G-protein-coupled P2Y receptors[3]. P2X receptors are found exclusively in eukaryotes, are expressed throughout the human body including the nervous, cardiovascular and immune systems, and are implicated in a wide range of physiological processes such as synaptic transmission, smooth muscle contraction, taste, nociception and inflammation[4,5]. Accordingly, P2X receptors hold great interest as new therapeutic targets for inflammatory, cardiovascular and neuronal disease[6].

P2X receptors are trimeric assemblies composed of seven distinct subunit subtypes ($P2X_{1-7}$)[7] that associate to form homomeric and heteromeric complexes[8,9]. All subunits share a common topology, with intracellular termini, two transmembrane domains and a large, glycosylated and disulphide-rich extracellular domain[2,10]. The extracellular domain contains binding sites for ATP, competitive antagonists and modulatory metal ions, whereas the transmembrane domains form a non-selective cation channel[11]. The gating properties of the ion channel by agonist vary markedly with receptor subtype, with the $P2X_2$, $P2X_4$ and $P2X_7$ homomeric channels showing slow desensitization and the $P2X_1$ and $P2X_3$ channels exhibiting rapid desensitization[7]. Although the P2X receptors are ostensibly non-selective cation permeable ion channels, several studies suggest that P2X receptors, especially after prolonged ATP application, are permeable to large organic cations such as *N*-methyl-D-glucamine (NMDG)[12,13]. The pharmacology of P2X receptor subtypes is also divergent, with the sole common feature being activation by ATP and variable yet marked allosteric modulation by metal ions, protons and lipophilic small molecules[14]. At present, we know little about the location and composition of the ATP-binding site, the conformational changes that ensue after ATP binding, and the nature of the open, ion-conducting pore of P2X receptors.

The zebrafish $P2X_4$ receptor[15] was recently crystallized in an apo, closed state and the resulting structure showed the chalice-shaped trimeric architecture of P2X receptors[16], defined the protein fold of the extracellular domains and illuminated the conformation of the closed ion channel pore. The closed state structure, in combination with previous mutational studies[17–22], suggested a location for three non-canonical and intersubunit ATP-binding sites. However, the absence of an experimental crystal structure in complex with ATP meant that the binding site for ATP, the mechanism of ATP-dependent gating and the nature of the open ion channel pore remained speculative. Here we report the crystal structures of the slowly desensitizing $P2X_4$ receptor in the presence and absence of ATP at 2.8 and 2.9 Å resolution, respectively. The ATP-bound structure reveals a previously unseen ATP-binding motif and an open pore conformation. Most importantly, a comparison of the two structures suggests how ATP binding is coupled to ion channel gating, thus providing the first, to our knowledge, structural insights into the agonist-induced activation mechanism of P2X receptors based on atomic resolution crystal structures.

Crystallization and structure determination

Initial crystals of the $\Delta P2X_4$-B–ATP complex, using the construct from the X-ray structure determination of the apo state[16], diffracted to only 7 Å resolution. We therefore screened new constructs of the $P2X_4$ receptor, in combination with further P2X receptor orthologues, by fluorescence-detection size-exclusion chromatography (FSEC)[23]. The deletion of several other residues from the carboxy terminus and the return of residue 51 to the native Phe yielded $\Delta P2X_4$-C—a construct that starts at Ser 28, ends at Lys 365 (ΔN27/ΔC24/N78K/N187R) and possesses a sharp and symmetric FSEC elution profile (Supplementary Fig. 1a). The $\Delta P2X_4$-C construct has similar ATP-binding and gating activities to the wild-type receptor[16], as judged by filter binding and two-electrode voltage clamp experiments, respectively (Supplementary Fig. 1b–f). Notably, $\Delta P2X_4$-C yields crystals in the presence of ATP that diffract X-rays to 2.8 Å resolution. The resulting structure was solved by molecular replacement using the previously solved $\Delta P2X_4$-B structure as a search probe, and refined

[1]Vollum Institute, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA. [2]Howard Hughes Medical Institute, Oregon Health and Science University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239, USA.

to good crystallographic statistics and stereochemistry (Supplementary Table 1 and Supplementary Figs 2–4).

We also measured X-ray diffraction data from apo state crystals of the $\Delta P2X_4$-B construct at a higher resolution (2.9 Å) than that obtained from crystals used in the initial structure determination at 3.1 Å resolution[16]. Analysis of electron density maps derived from the higher resolution apo and ATP-bound data sets allowed us to correct an incorrect registration of residues 88–97, by one residue, present in the initial apo state structure (Supplementary Fig. 5 and Supplementary Table 1). Because the new $\Delta P2X_4$-B structure (termed $\Delta P2X_4$-B_2) is more accurate than the original structure (termed $\Delta P2X4$-B_1), we use the former in comparisons with the ATP-bound state.

Architecture

The ATP-bound $\Delta P2X_4$-C receptor adopts a chalice-shaped, homotrimetric architecture consisting of a large hydrophilic and glycosylated extracellular domain, a transmembrane domain composed of 6 α-helices and short intracellular amino and carboxy termini (Fig. 1). Each subunit resembles the shape of a dolphin, with the transmembrane helices and the extracellular region akin to the flukes and the body, respectively (Fig. 2a). The protein fold and overall structure of an ATP-complexed $\Delta P2X_4$-C subunit is similar to that of an apo $\Delta P2X_4$-B_2 subunit (Figs 1 and 2a), as illustrated by a root mean squared deviation (r.m.s.d.) of 1.8 Å for the Cα atom position after superposition[16]. Superposition of the ATP-bound and apo trimers, however, yields r.m.s.d. values of 3.2 Å for the Cα atoms, thus demonstrating that substantial conformational changes are associated with the binding of ATP, with some of the largest differences found at and adjacent to the ATP-binding sites in the extracellular domain and within the ion-conducting transmembrane domain (Fig. 2b). In the context of the trimeric receptor, superposition of the apo, closed state with the ATP-bound state demonstrates that the upper body domain

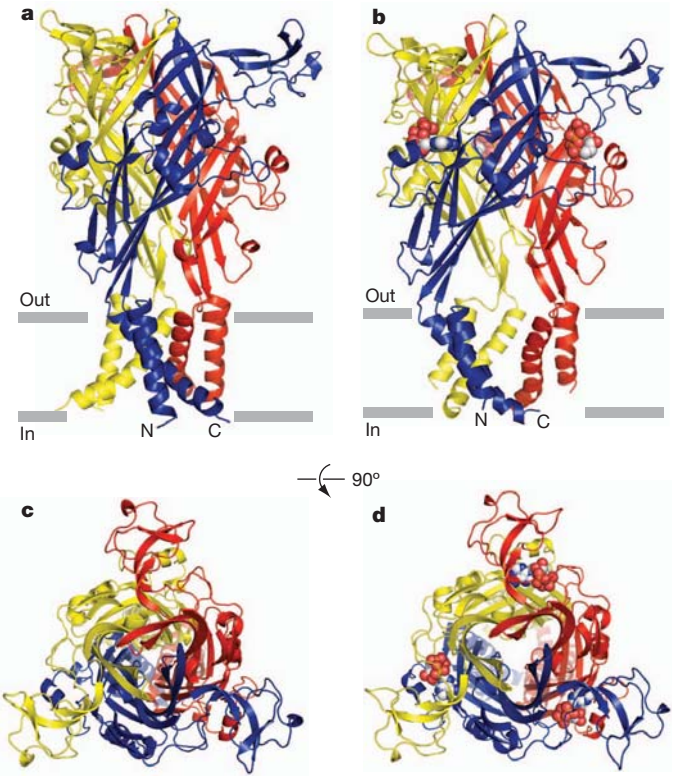


Figure 1 | The architectures of zebrafish $P2X_4$. **a**, **b**, Zebrafish $\Delta P2X_4$-B_2 (**a**) and $\Delta P2X_4$-C (**b**) trimer structures viewed parallel to the membrane. Each subunit is shown in a different colour. ATP is shown in sphere representation. **c**, **d**, Zebrafish $\Delta P2X_4$-B_2 (**c**) and $\Delta P2X_4$-C (**d**) trimer structures viewed from the extracellular side.

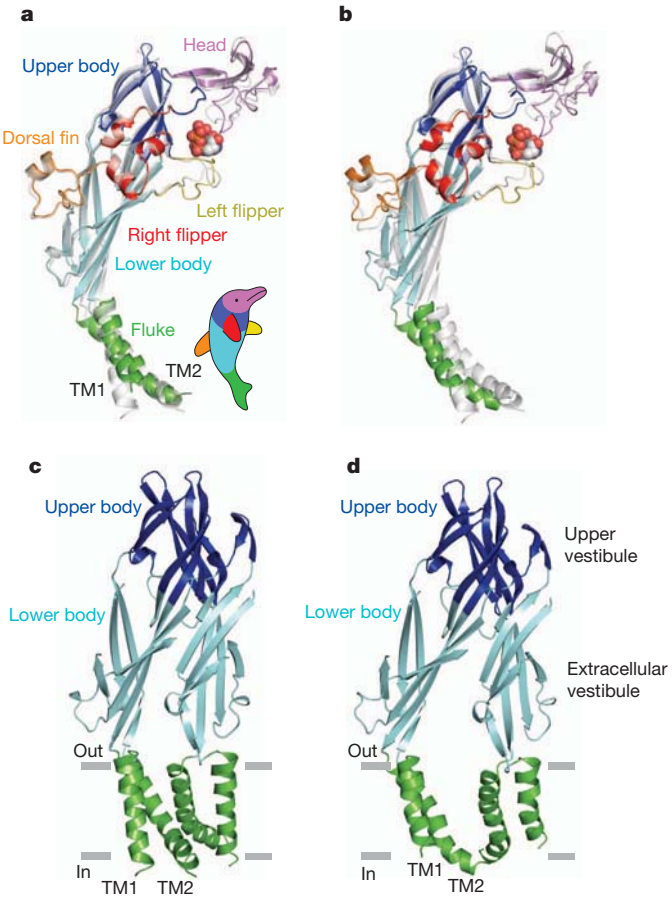


Figure 2 | Analysis of conformational difference between $\Delta P2X_4$-C and $\Delta P2X_4$-B_2. Each structural feature of the dolphin-shaped $P2X_4$ subunit is coloured differently. **a**, **b**, Superpositions between $\Delta P2X_4$-C and $\Delta P2X_4$-B_2 (grey) using Cα positions of the protomers (**a**) and trimers (**b**). **c**, **d**, The transmembrane and body domains of $\Delta P2X_4$-B_2 (**c**) and $\Delta P2X_4$-C (**d**). Only two subunits in the foreground are shown.

does not undergo substantial conformational changes after ATP binding, thus suggesting that the upper body domain is a relatively rigid 'scaffold' or 'brace' (Fig. 2c, d). The lower body domain, by contrast, does not superimpose well after comparison of the apo and ATP-bound states because of an outward 'flexing' of the body domain (Fig. 2c, d). This conformational difference between the two states expands the lower region of the extracellular vestibule, increasing the separation between the Cα atoms of Asp 59 residues on adjacent subunits from 15.0 Å in the apo state to 25.5 Å in the ATP-bound form. The notable movement of the lower body domains, in turn, directly expands the transmembrane helices (Fig. 2c, d). Overall, analysis of the ATP-bound structure, in combination with comparison to the apo structure, allows us to define the agonist-binding site and to propose a mechanism by which agonist binding leads to ion channel gating in P2X receptors.

ATP-binding site

Inspection of electron density maps derived from $\Delta P2X_4$-C–ATP cocrystals immediately showed a prominent feature consistent with the molecular composition of an ATP molecule (Fig. 3 and Supplementary Figs 6 and 7). Application of the crystallographic symmetry inherent in the *R*32 space group of this crystal form yields three equivalent ATP-binding sites at each of the three pairs of subunit interfaces in the trimeric receptor, sites located ~40 Å from the extracellular boundary of the transmembrane domain. The intersubunit ATP-binding pocket[19,24], lined with several positively charged residues

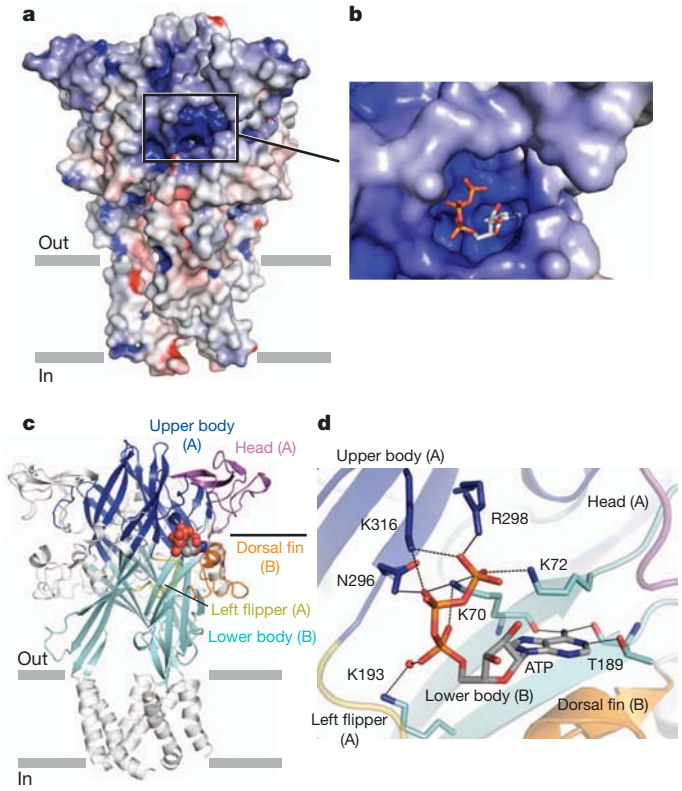


Figure 3 | ATP-binding site. **a**, **b**, An electrostatic potential surface of $\Delta P2X_4$-C (**a**), contoured from −10 kT (red) to +10 kT (blue) (dielectric constant: 80), and its close-up view (**b**). **c**, The regions forming the ATP-binding pocket are coloured as in Fig. 2a. The ATP molecule is shown in sphere representation. **d**, Close-up view of the ATP-binding site. The oxygen atom from the glycerol molecule is shown in sphere representation. Black dashed lines indicate hydrogen bonding (<3.3 Å).

(Figs 1b and 3a, b), is cradled by the head domain (chain A), upper body (chain A), lower body (chain B), left flipper (chain A) and the dorsal fin (chain B) (Fig. 3c). ATP is recognized by the upper (chain A) and lower (chain B) body domains through extensive hydrophilic interactions (Fig. 3d) derived from an underlying protein fold that is, to the best of our knowledge, different from any other known ATP-binding motif. The head domain, left flipper and dorsal fin participate in several further direct contacts with ATP (Fig. 3d and Supplementary Fig. 7), consistent with their essential participation in the ATP-binding site (Fig. 2).

The bound ATP molecule adopts a U-shaped structure, with the β- and γ-phosphates folded towards the adenine ring and the base–sugar constellation in an 'anti' conformation (Fig. 3d)—a conformation previously observed in class II aminoacyl transfer RNA synthetases[25]. In the $\Delta P2X_4$-C complex with ATP, the U-shaped conformation of negatively charged phosphate groups participate in salt bridge and hydrogen bonding interactions, with a cluster of highly conserved basic and polar residues emanating from the two subunits that together form the agonist-binding pocket (Fig. 3d). Using the ATP molecule bound at the agonist site between the A and B subunit as an example, Lys 70 (chain B) occupies a crucial site because its ammonium group resides at the centre of the triphosphate 'U', forming interactions with oxygen atoms on the α, β and γ phosphate groups. Asn 296 and Lys 316 from chain A mediate further contacts with β-phosphate groups, whereas Lys 72 (chain B), Arg 298 (chain A) and Lys 316 (chain A) participate in interactions with the γ-phosphate (Fig. 3d).

The extensive intersubunit interactions with the β- or γ-phosphate groups provide a plausible explanation for why ADP and AMP have very weak or no effect on the activation of P2X receptors[26]. Yet the triphosphate moiety is not entirely buried in the agonist-binding pocket and the β and γ groups are partially exposed to solvent, consistent with the observation that diadenosine polyphosphates are full or partial agonists at rat P2X receptors[27,28]. Interactions between the receptor and ATP are also mediated by solvent molecules, and in the crystal structure a glycerol molecule bridges interactions between Lys 193 and the α-phosphate (Fig. 3d and Supplementary Fig. 8a) and a second glycerol molecule participates in interactions with the β- and γ-phosphates (Supplementary Fig. 8b). Under physiological conditions we suggest that several water molecules, instead of the glycerol molecules, occupy these sites.

The adenine base of ATP is deeply buried in the ATP-binding pocket (Fig. 3b), and is recognized by three hydrogen bonds with the side chain of Thr 189 and the main-chain carbonyl oxygen atoms of Lys 70 and Thr 189 in the lower body (Fig. 3c, d and Supplementary Fig. 7a). All these residues are strictly conserved[16] and have been implicated in the ATP-dependent gating of P2X receptors[17–22] and in the adenine base recognition[29]. There are also hydrophobic interactions between the adenine base and Leu 191 in the lower body and Ile 232 in the dorsal fin (Supplementary Fig. 7b). The hydrophobicity of these residues is conserved[16], and Leu 186 in rat $P2X_2$, corresponding to Leu 191 in zebrafish $P2X_4$, is suggested to be involved in the recognition of the adenine base[30].

The ribose ring of ATP is recognized only by Leu 217 in the dorsal fin through hydrophobic interactions (Supplementary Fig. 7b), and the O2 and O3 atoms of the ribose ring are solvent-accessible (Fig. 3b). Consistent with the solvent exposure of the ribose O2 and O3 atoms, ribose-modified ATP analogues can activate or inhibit P2X receptors[31,32].

Base specificity

P2X receptors preferentially recognize ATP, whereas CTP has at most a weak effect on receptor activity[33,34] and GTP and UTP do not activate P2X receptors[26]. To understand the mechanism by which P2X receptors achieve this specificity, we superimposed GTP, CTP and UTP onto the ATP molecule bound to the $\Delta P2X_4$-C structure and measured the binding of the nucleotide triphosphate molecules with a competition assay using ^{3}H-ATP (Supplementary Fig. 9).

CTP and ATP contain similar 'amidine' functional groups within their base ring structures, and on the basis of our superposition, CTP can form hydrogen bonds between the base N4 atom and the carbonyl oxygen atoms of Lys 70 and possibly Thr 189 (Supplementary Fig. 9b). Because the base of CTP is smaller than that of ATP, however, the N3 atom is too far from the side chain of Thr 189 to form a hydrogen bond. In addition, the smaller cytidine base does not fill the agonist-binding pocket and the resulting cavity probably further diminishes the extent to which CTP can bind to and activate P2X receptors. GTP and UTP, in contrast to ATP and CTP, possess nearly reciprocal hydrogen bonding groups on their base rings (Supplementary Fig. 9a–d) and are therefore unable to form favourable hydrogen bonding interactions with carbonyl oxygen atoms of Lys 70 and Thr 189, thus providing a chemical explanation for why GTP and UTP bind with low affinity to P2X receptors.

Open pore conformation

The $\Delta P2X_4$-C structure reveals an uninterrupted, continuous transmembrane pore (Fig. 4). The pore is lined primarily by TM2 with residues Leu 340, Ala 344, Ala 347, Leu 351 and Ile 355 exposed to the putative ion permeation pathway (Fig. 4d), an observation consistent with cysteine-accessibility studies[35–39]. In comparison to the $\Delta P2X_4$-B_2 apo state structure, in which Leu 340 and Ala 347 define the extracellular and intracellular boundaries of the closed ion channel gate, respectively, the most narrow diameters of the ion conductive pathway in the $\Delta P2X_4$-C structure are at Ala 347 and Leu 351 (~7 Å), observations in accord with experiments on the permeability of P2X receptors to organic cations[40,41] (Fig. 4d, e and Supplementary

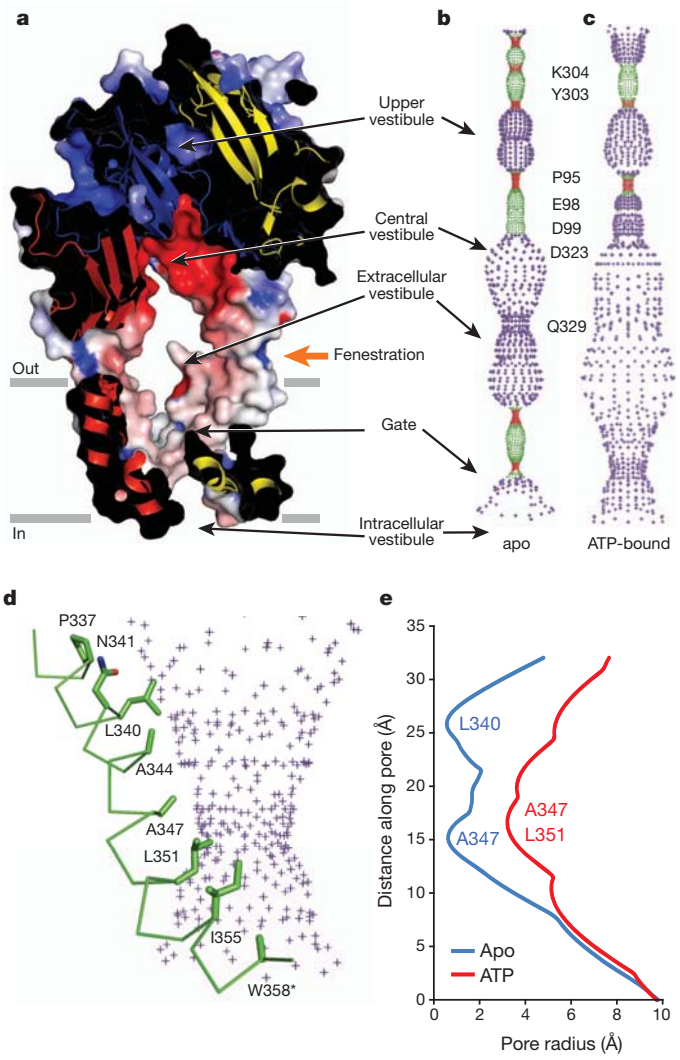


Figure 4 | The transmembrane pore. **a**, A section of an electrostatic potential surface of $\Delta P2X_4$-C, contoured as in Fig. 3a. Pore-lining surfaces of $\Delta P2X_4$-B_2 (**b**) and $\Delta P2X_4$-C (**c**). Each colour indicates a different radius range from the pore centre (red: <1.15 Å, green: 1.15–2.3 Å, and purple: >2.3 Å). **d**, Pore-lining residues of $\Delta P2X_4$-C shown in stick representation with the pore-lining surface. **e**, Pore radius for $\Delta P2X_4$-C and $\Delta P2X_4$-B_2 along the pore centre axis.

Fig. 10). Thus, we propose that this ATP-bound structure represents an activated, open channel conformation.

$P2X_4$ receptors are well known for undergoing 'pore dilation' after prolonged application of ATP, a process defined by increased permeability to large organic cations such as NMDG, with a mean diameter of 7.3 Å[12,13,42]. To probe the functional properties of the $\Delta P2X_4$-C ion channel pore, we carried out two-electrode voltage clamp studies, examining whether prolonged application of ATP leads to pore dilation. After a 5-min application of saturating ATP, we find that the evoked current remains constant, suggesting that the $\Delta P2X_4$-C construct represents a non-pore-dilated open state (Supplementary Fig. 1g, h).

Interestingly, there are no hydrophilic residues lining the middle of the pore (Fig. 4d). Therefore, water molecules coordinated to permeating cations probably interact with main-chain carbonyl oxygen and nitrogen atoms. For the rat $P2X_2$ receptor, however, a threonine implicated in ion selectivity is at the position equivalent to Ala 347 in zebrafish $P2X_4$ (Fig. 4d, e)[16], near the constriction region in the pore, thus suggesting that this protein side chain interacts directly with permeant hydrated cations[43,44].

Ion access to the pore

The previous zebrafish $\Delta P2X_4$-B_1 structure suggested two possible pathways by which cations might access the ion channel: a central pathway, along the three-fold axis of symmetry, and a lateral pathway through the fenestrations 'above' the ion channel pore (Fig. 4a–c)[16]. In the open-state, ATP-bound $\Delta P2X_4$-C structure, the pathway along the three-fold axis of symmetry is too small to allow for ion permeation, whereas the lateral fenestrations in the extracellular vestibule of $\Delta P2X_4$-C are wide open (Fig. 4a–c)[16]. Therefore, the lateral fenestrations are the pathways by which hydrated ions enter and exit the receptor, in agreement with recent cysteine-accessibility, cysteine-based cross-linking and computational studies[45,46]. Once ions pass through the fenestrations, the highly acidic central vestibule attracts cations and repels anions, thus concentrating cations close to the entrance of the ion channel pore[46].

Structural transition in the transmembrane domain

A structural comparison of the transmembrane regions of the closed and open states shows that the transmembrane helices rearrange in an iris-like movement in going from the closed to the open state (Fig. 5). Relative to the closed state structure, transmembrane domain 1 (TM1) and TM2 rotate by ~10° and ~55° anticlockwise about a pore-centre axis perpendicular to the membrane plane, and increase their tilt angle by ~8° and ~2° about an axis parallel to the membrane plane, respectively (Fig. 5a, b and Supplementary Movie 1). The consequence of the iris-like movement of TM1 and TM2 helices is that the

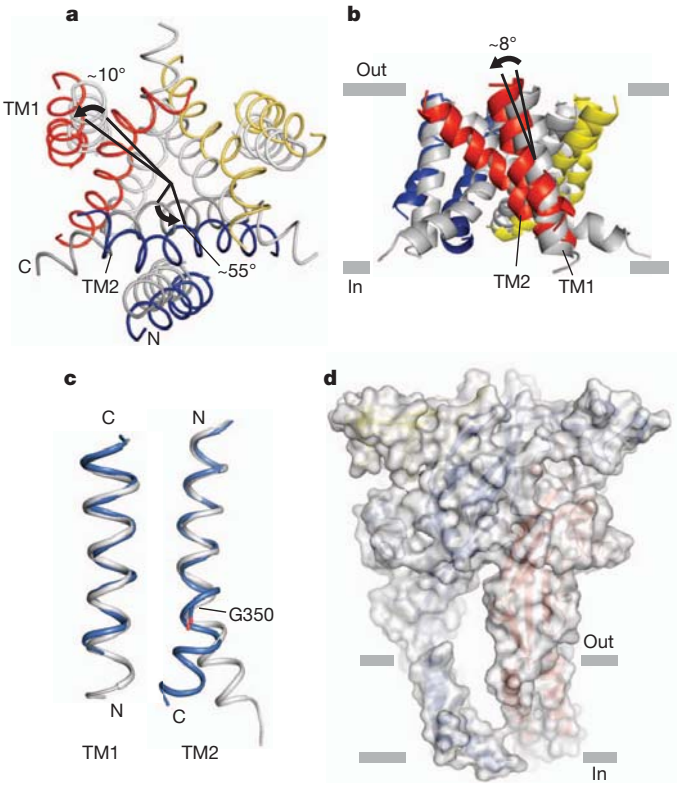


Figure 5 | Structural transitions in the transmembrane domain. **a**, **b**, The transmembrane region of $\Delta P2X_4$-C and $\Delta P2X_4$-B_2 (grey), viewed from the intracellular side (**a**) and parallel to the membrane (**b**). $\Delta P2X_4$-B_2 is superimposed on $\Delta P2X_4$-C using Cα positions of the trimer. The black arrows and bars denote the rotation of the transmembrane helices (**a**) and the orientation of the transmembrane helices (**b**), respectively. **c**, Close-up view of the transmembrane helices. Transmembrane helices of $\Delta P2X_4$-B_2 (grey) are superimposed on those of $\Delta P2X_4$-C using Cα positions of residue 36–55 for TM1 and residue 334–349 for TM2, respectively. Gly 350 is shown in stick representation. **d**, A surface model of the $\Delta P2X_4$-C trimer with the cartoon representation inside.

helices move away from the central axis by ~3 Å to expand an ion conductive pore (Supplementary Fig. 10 and Supplementary Movie 1).

The transition from closed to open pore greatly alters intra- and intersubunit interactions between the transmembrane helices (Supplementary Fig. 11). In the apo state of the receptor, several interactions between TM2 helices that include contacts between Leu 340, Leu 346 and Ala 347 stabilize the closed conformation of the pore (Supplementary Fig. 11c). In the ATP-bound open state, these interactions are ruptured as the transmembrane helices move away from the three-fold axis. Accompanying the radial movement of the transmembrane helices from the three-fold axis is a kinking of the TM2 helix that allows for new contacts to form between subunits involving Leu 346 and Ile 355, stabilizing the wide-opening pore conformation (Supplementary Fig. 11d). The kink in TM2 is localized to Gly 350 (Fig. 5c, Supplementary Fig. 11 and Supplementary Movie 1), a conserved residue previously suggested to function as a gating hinge[12,47]. Gly 350 is associated with weak electron density (Supplementary Fig. 4b), an observation consistent with its role as a flexible hinge.

The movement of the transmembrane helices away from the three-fold axis creates notable 'gaps' between the transmembrane helices of adjacent subunits, voids that must be occupied by lipid molecules when the channel resides in its native membrane environment (Fig. 5d and Supplementary Fig. 11c, d). Interestingly, residues lining the gaps between subunits and near the kink in TM2 include amino acids implicated in interactions with ivermectin[48,49], a positive allosteric modulator of $P2X_4$ receptors (Supplementary Fig. 12)[50]. We therefore suggest that ivermectin, and perhaps endogenous lipids, may occupy the gap between transmembrane helices in the open state, and by so doing allosterically modulate the activity of the channel.

Mechanism of activation

A structural comparison of the extracellular regions of the apo, closed and agonist-bound open states shows how ATP binding leads to channel activation. First, at the ATP-binding site, within the intra-subunit cleft, ATP promotes cleft closure between the head and dorsal fin domains, causing the movement of the dorsal fin domain 'up' towards the head domain to accommodate ATP via hydrophobic interactions, whereas ATP pushes out the left flipper from the ATP-binding pocket (Fig. 6, Supplementary Figs 13, 14 and Supplementary Movie 2). Because the dorsal fin and left flipper are structurally coupled to the lower body domain (Supplementary Figs 13, 14 and Supplementary Movie 2), there is a concomitant outward flexing of the lower body domain in the ATP-bound state that substantially expands the extracellular vestibule, increasing the separation by ~10 Å (Figs 2c, d and 6c, d). In the flexing or 'lever arm' motion of the lower body domains, the upper body domains of each subunit largely behave as a rigid body, or brace (Supplementary Fig. 15), and subunits rotate by ~8° around a rotation axis located in the upper body (Fig. 6a). Finally, the lower body domains are directly coupled to TM1 and TM2 and thus their outward flexing directly promotes the opening of the ion channel pore by causing the transmembrane helices to 'expand' in an iris-like motion (Figs 2c, d and 6c, d and Supplementary Movie 2).

In the context of this structure-based mechanism we speculate that competitive antagonists, such as 2′,3′-O-(2,4,6-trinitrophenyl)–ATP (TNP–ATP)[32], antagonize the receptor by binding to the ATP site while blocking dorsal fin closure and subsequent ion channel gating because of the steric bulk of the trinitrophenyl moieties (Supplementary Fig. 16).

Conclusion

The crystal structure of the $P2X_4$-C receptor in complex with ATP shows, to the best of our knowledge, that P2X receptors contain an ATP-binding motif not previously seen before. Agonist binding promotes cleft closure in an intersubunit-binding site, resulting in the flexing of the lower body domain that in turn directly expands the region of the receptor proximal to the ion channel pore, causing an iris-like opening of an ion conductive pathway. After transition to the open state, potential ion pathways in the extracellular domain along the three-fold axis remain occluded by protein, and cations instead gain access to the ion channel by way of three lateral fenestrations. The open state of the receptor is characterized by few subunit–subunit contacts within the transmembrane domains and these relatively large gaps are probably occupied by lipids and allosteric modulators such as ivermectin. Sparse contacts between transmembrane domains suggest that the ion channel domain may adopt several conformations, a structural observation consistent with a multiplicity of pore selectivity states deduced from electrophysiological studies of P2X receptors. Taken together, this work illustrates how the binding of ATP activates P2X receptors and initiates ionotropic purinergic signalling.

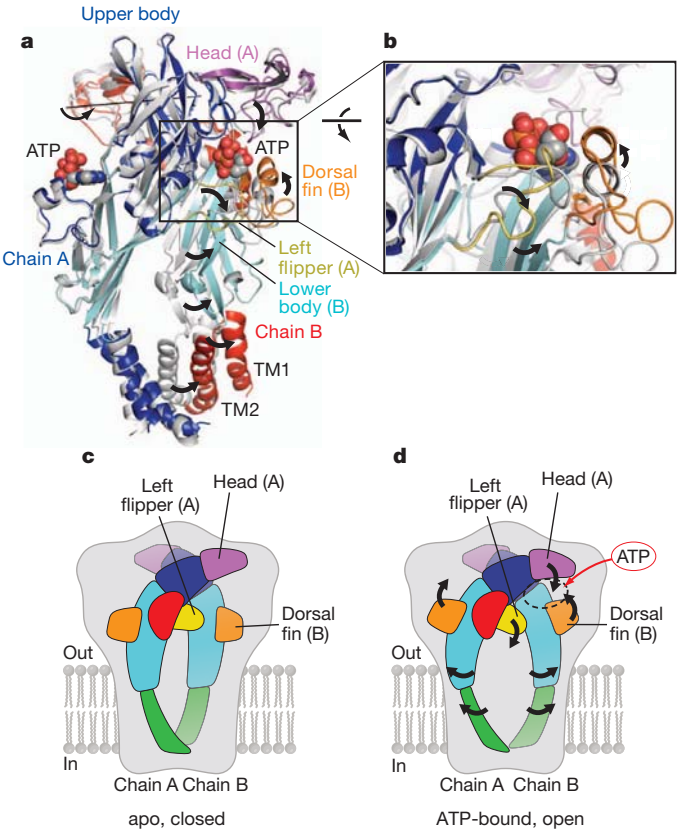


Figure 6 | Mechanism of activation. **a**, A $\Delta P2X_4$-B_2 subunit (grey) is superimposed on $\Delta P2X_4$-C using Cα positions of protomer A. Only two subunits in the foreground are shown. The rotation axis describes the superposition of the apo $\Delta P2X_4$-B_2 B subunit onto the ATP-bound $\Delta P2X_4$-C B protomer. **b**, Close-up view of the conformational changes resulting from ATP binding. **c**, **d**, A cartoon model of the ATP-dependent activation mechanism. The black arrows denote the movement from the apo closed state (**c**) to the ATP-bound open state (**d**).

METHODS SUMMARY

The zebrafish $\Delta P2X_4$-C and $\Delta P2X_4$-B proteins were expressed as N-terminal octa-histidine- enhanced green fluorescent protein (EGFP) fusions in baculovirus-infected Sf9 cells, and were purified as described previously[16]. For samples used in crystallization, 1 mM ATP and 1 mM $GdCl_3$ were added to purified $\Delta P2X_4$-C and $\Delta P2X_4$-B, respectively. Apo state $\Delta P2X_4$-B_2 crystals were grown at 4 °C by vapour diffusion using a reservoir solution containing 18–22% PEG 3350, 100 mM $MgCl_2$, 2 M NaCl and 0.1 M imidazole, pH 6.5. For ATP-bound $\Delta P2X_4$-C crystals, growth occurred at 4 °C by vapour diffusion with a reservoir solution containing 20–26% PEG 2000, 300 mM $Mg(NO_3)_2$ and 100 mM Tris, pH 8.0. Diffraction data were processed and the structures were solved by molecular replacement. The resulting models were then subjected to iterative cycles of manual adjustment and crystallographic refinement. The functional

properties of the $\Delta P2X_4$-C construct were examined by two-electrode voltage clamp experiments and by [^{3}H]-ATP saturation binding assays.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 November 2011; accepted 4 March 2012.
Published online 25 April; corrected 9 May 2012 (see full-text HTML version for details).

1. Burnstock, G. Purinergic nerves. *Pharmacol. Rev.* **24,** 509–581 (1972).
2. Valera, S. *et al.* A new class of ligand-gated ion channel defined by P_{2x} receptor for extracellular ATP. *Nature* **371,** 516–519 (1994).
3. Webb, T. E. *et al.* Cloning and functional expression of a brain G-protein-coupled ATP receptor. *FEBS Lett.* **324,** 219–225 (1993).
4. Surprenant, A. & North, R. A. Signaling at purinergic P2X receptors. *Annu. Rev. Physiol.* **71,** 333–359 (2009).
5. Burnstock, G. & Kennedy, C. P2X receptors in health and disease. *Adv. Pharmacol.* **61,** 333–372 (2011).
6. Coddou, C., Yan, Z., Obsil, T., Huidobro-Toro, J. P. & Stojikovic, S. S. Activation and regulation of purinergic P2X receptor channels. *Pharmacol. Rev.* **63,** 641–683 (2011).
7. North, R. A. Molecular physiology of P2X receptors. *Physiol. Rev.* **82,** 1013–1067 (2002).
8. Nicke, A. *et al.* P2X1 and P2X3 receptors form stable trimers: a novel structural motif of ligand-gated ion channels. *EMBO J.* **17,** 3016–3028 (1998).
9. Aschrafi, A., Sadtler, S., Niculescu, C., Rettinger, J. & Schmalzing, G. Trimeric architecture of homomeric $P2X_2$ and heteromeric $P2X_{1+2}$ receptor subtypes. *J. Mol. Biol.* **342,** 333–343 (2004).
10. Brake, A. J., Wagenbach, M. J. & Julius, D. New structural motif for ligand-gated ion channels defined by an ionotropic ATP receptor. *Nature* **371,** 519–523 (1994).
11. Browne, L. E., Jiang, L. H. & North, R. A. New structure enlivens interest in P2X receptors. *Trends Pharmacol. Sci.* **31,** 229–237 (2010).
12. Khakh, B. S., Bao, X. R., Labarca, C. & Lester, H. A. Neuronal P2X transmitter-gated cation channels change their ion selectivity in seconds. *Nature Neurosci.* **2,** 322–330 (1999).
13. Virginio, C., MacKenzie, A., Rassendren, F. A., North, R. A. & Surprenant, A. Pore dilation of neuronal P2X receptor channels. *Nature Neurosci.* **2,** 315–321 (1999).
14. Jarvis, M. F. & Khakh, B. S. ATP-gated P2X cation-channels. *Neuropharmacology* **56,** 208–215 (2009).
15. Kucenas, S., Li, Z., Cox, J. A., Egan, T. M. & Voigt, M. M. Molecular characterization of the zebrafish P2X receptor subunit gene family. *Neuroscience* **121,** 935–945 (2003).
16. Kawate, T., Michel, J. C., Birdsong, W. T. & Gouaux, E. Crystal structure of the ATP-gated $P2X_4$ ion channel in the closed state. *Nature* **460,** 592–598 (2009).
17. Jiang, L. H., Rassendren, F., Surprenant, A. & North, R. A. Identification of amino acid reisdues contributing to the ATP-binding site of a purinergic P2X receptor. *J. Biol. Chem.* **275,** 34190–34196 (2000).
18. Ennion, S., Hagan, S. & Evans, R. J. The role of positively charged amino acids in ATP recognition by human $P2X_1$ receptors. *J. Biol. Chem.* **275,** 29361–29367 (2000).
19. Marquez-Klaka, B., Rettinger, J., Bhargava, Y., Eisele, T. & Nicke, A. Identification of an intersubunit cross-link between substituted cysteine residues located in the putative ATP binding site of the $P2X_1$ receptor. *J. Neurosci.* **27,** 1456–1466 (2007).
20. Roberts, J. A. & Evans, R. J. Cysteine substitution mutants give structural insight and identify ATP binding and activation sites at P2X receptors. *J. Neurosci.* **27,** 4072–4082 (2007).
21. Roberts, J. A. *et al.* Cysteine substitution mutagenesis and the effects of methanethiosulfonate reagents at $P2X_2$ and $P2X_4$ receptors support a core common mode of ATP action at P2X receptors. *J. Biol. Chem.* **283,** 20126–20136 (2008).
22. Bodnar, M. *et al.* Amino acid residues constituting the agonist binding site of the human P2X3 receptor. *J. Biol. Chem.* **286,** 2739–2749 (2011).
23. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14,** 673–681 (2006).
24. Wilkinson, W. J., Jiang, L. H., Surprenant, A. & North, R. A. Role of ectodomain lysines in the subunits of the heteromeric $P2X_{2/3}$ receptor. *Mol. Pharmacol.* **70,** 1159–1163 (2006).
25. Cavarelli, J. *et al.* The active site of yeast aspartyl-tRNA synthetase: structural and functional aspects of the aminoacylation reaction. *EMBO J.* **13,** 327–337 (1994).
26. Gever, J. R., Cockayne, D. A., Dillon, M. P., Burnstock, G. & Ford, A. P. D. W. Pharmacology of P2X channels. *Pflugers Arch.* **452,** 513–537 (2006).
27. Evans, R. J. *et al.* Pharmacological characterization of heterologously expressed ATP-gated cation channels (P2x purinoceptors). *Mol. Pharmacol.* **48,** 178–183 (1995).
28. Wildman, S. S., Brown, S. G., King, B. F. & Burnstock, G. Selectivity of diadenosine polyphosphates for rat P2X receptor subunits. *Eur. J. Pharmacol.* **367,** 119–123 (1999).
29. Roberts, J. A., Valente, M., Allsopp, R. C., Watt, D. & Evans, R. J. Contribution of the region Glu181 to Val200 of the extracellular loop of the human P2X1 receptor to agonist binding and gating revealed using cysteine scanning mutagenesis. *J. Neurochem.* **109,** 1042–1052 (2009).
30. Jiang, R. *et al.* Agonist trapped in ATP-binding sites of the P2X2 receptor. *Proc. Natl Acad. Sci. USA* **108,** 9066–9071 (2011).
31. Bianchi, B. R. *et al.* Pharmacological characterization of recombinant human and rat P2X receptor subtypes. *Eur. J. Pharmacol.* **376,** 127–138 (1999).
32. Virginio, C., Robertson, G., Surprenant, A. & North, R. A. Trinitrophenyl-substituted nucleotides are potent antagonists selective for $P2X_1$, $P2X_3$, and heteromeric $P2X_{2/3}$ receptors. *Mol. Pharmacol.* **53,** 969–973 (1998).
33. Soto, F. *et al.* P2X4: an ATP-activated ionotropic receptor cloned from rat brain. *Proc. Natl Acad. Sci. USA* **93,** 3684–3688 (1996).
34. Garcia-Guzman, M., Soto, F., Gomez-Hernandez, J. M., Lund, P. E. & Stühmer, W. Characterization of recombinant human P2X4 receptor reveals pharmacological differences to the rat homologue. *Mol. Pharmacol.* **51,** 109–118 (1997).
35. Egan, T. M., Haines, W. R. & Voigt, M. M. A domain contributing to the ion channel of ATP-gated $P2X_2$ receptors identified by the substituted cysteine accessibility method. *J. Neurosci.* **18,** 2350–2359 (1998).
36. Rassendren, F., Buell, G., Newbolt, A., North, R. A. & Surprenant, A. Identification of amino acid residues contributing to the pore of a P2X receptor. *EMBO J.* **16,** 3446–3454 (1997).
37. Li, M., Chang, T. H., Silberberg, S. D. & Swartz, K. J. Gating the pore of P2X receptor channels. *Nature Neurosci.* **11,** 883–887 (2008).
38. Li, M., Kawate, T., Silberberg, S. D. & Swartz, K. J. Pore-opening mechanism in trimeric P2X receptor channels. *Nature Commun.* **1,** 1–7 (2010).
39. Kracun, S., Chaptal, V., Abramson, J. & Khakh, B. S. Gated access to the pore of a P2X receptor: structural implications for closed-open transitions. *J. Biol. Chem.* **285,** 10110–10121 (2010).
40. Evans, R. J. *et al.* Ionic permeability of, and divalent cation effects on, two ATP-gated cation channels (P2X receptors) expressed in mammalian cells. *J. Physiol. (Lond.)* **497,** 413–422 (1996).
41. Liu, D. M. & Adams, D. J. Ionic selectivity of native ATP-activated (P2X) receptor channels in dissociated neurones from rat parasympathetic ganglia. *J. Physiol. (Lond.)* **534,** 423–435 (2001).
42. Villarroel, A., Burnashev, N. & Sakmann, B. Dimensions of the narrow portion of a recombinant NMDA receptor channel. *Biophys. J.* **68,** 866–875 (1995).
43. Migita, K., Haines, W. R., Voigt, M. M. & Egan, T. M. Polar residues of the second transmembrane domain influence cation permeability of the ATP-gated $P2X_2$ receptor. *J. Biol. Chem.* **276,** 30934–30941 (2001).
44. Browne, L. E. *et al.* P2X receptor channels show threefold symmetry in ionic charge selectivity and unitary conductance. *Nature Neurosci.* **14,** 17–18 (2011).
45. Kawate, T., Robertson, J. L., Li, M., Silberberg, S. D. & Swartz, K. J. Ion access pathway to the transmembrane pore in P2X receptor channels. *J. Gen. Physiol.* **137,** 579–590 (2011).
46. Samways, D. S., Khakh, B. S., Dutertre, S. & Egan, T. M. Preferential use of unobstructed lateral portals as the access route to the pore of human ATP-gated ion channels (P2X receptors). *Proc. Natl Acad. Sci. USA* **108,** 13800–13805 (2011).
47. Fujiwara, Y., Keceli, B., Nakajo, K. & Kubo, Y. Voltage- and [ATP]-dependent gating of the $P2X_2$ ATP receptor channel. *J. Gen. Physiol.* **133,** 93–109 (2009).
48. Jelínková, I. *et al.* Identification of $P2X_4$ receptor-specific residues contributing to the ivermectin effects on channel activation. *Biochem. Biophys. Res. Commun.* **349,** 619–625 (2006).
49. Silberberg, S. D., Li, M. & Swartz, K. J. Ivermectin interaction with transmembrane helices reveals widespread rearrangements during opening of P2X receptor channels. *Neuron* **54,** 263–274 (2007).
50. Khakh, B. S., Proctor, W. R., Dunwiddie, T. V., Labarca, C. & Lester, H. A. Allosteric control of gating and kinetics at $P2X_4$ receptor channels. *J. Neurosci.* **19,** 7289–7299 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Alexander and D. C. Dawson for providing *Xenopus* oocytes, K. L. Dürr and T. Friedrich for providing the pCNA3.1x vector, L. Vaskalis for assistance with figure illustrations, K. J. Swartz and M. P. Kavanaugh for advice related to the oocyte experiments, and Gouaux laboratory members for discussions. We are also grateful to the staff at the Advanced Photon Source beamline 24-ID-C for help with X-ray data collection. This work was supported by a Japan Society for the Promotion of Science Postdoctoral Fellowship for Research Abroad (M.H.), by the American Asthma Foundation (E.G.) and the NIH (E.G.). E.G. is an investigator with the Howard Hughes Medical Institute.

Author Contributions M.H. and E.G. contributed to all aspects of the project.

Author Information The coordinates and structure factors for the zebrafish apo $\Delta P2X4\text{-}B_2$ and ATP-bound $\Delta P2X4$-C have been deposited in the Protein Data Bank under the accession codes 4DW0 and 4DW1, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.G. (gouauxe@ohsu.edu).

METHODS

Expression and purification. The biochemically well-behaved and functionally active construct of zebrafish $P2X_{4.1}$ ($\Delta P2X_4$-C) was discovered by examining C-terminal deletion constructs of the zebrafish receptor as well as several additional P2X receptors from other organisms. These receptor candidates were screened by expression in Sf9 insect or human embryonic kidney cells and analysed by FSEC[23]. The zebrafish $\Delta P2X_4$-C and $\Delta P2X_4$-B proteins were expressed as N-terminal EGFP fusions with an octa-histidine affinity tag in baculovirus-infected Sf9 cells and were purified as described[16]. For crystallization, 1 mM ATP and 1 mM $GdCl_3$ were added to purified $\Delta P2X_4$-C and $\Delta P2X_4$-B from 100 mM stock solutions, respectively.

Crystallization. The $\Delta P2X_4$-B_2 crystals were obtained at 4 °C in two weeks by vapour diffusion by mixing 1:1 or 2:1 (v/v) ratios of protein and a reservoir solution containing 18–22% PEG 3350, 100 mM $MgCl_2$, 2 M NaCl and 0.1 M imidazole, pH 6.5. Crystals were collected in a harvest solution containing 20% PEG 3350, 100 mM $MgCl_2$, 1.3 M NaCl, 0.1 M imidazole, pH 6.5, 16% glycerol, 0.5 mM 1 mM *n*-dodecyl β-D-maltoside (DDM) and 1 mM TNP-ATP, incubated at 4 °C overnight, and cryoprotected by adding glycerol in 4% steps (final 24%). To minimize the occupancy of Gd in the receptor structure, $GdCl_3$ was excluded from harvest and cryoprotection solutions. The $\Delta P2X_4$-C crystals were grown at 4 °C in three days by vapour diffusion by mixing 1:1 or 2:1 ratios of protein and a reservoir solution containing 20–26% PEG 2000, 300 mM $Mg(NO_3)_2$ and 100 mM Tris, pH 8.0. Crystals were collected in a solution containing 25% PEG 2000, 300 mM $Mg(NO_3)_2$, 100 mM Tris, pH 8.0, 15% glycerol, 0.5 mM DDM and 1 mM ATP, and cryoprotected by adding glycerol in 5% steps (final 25%). Crystals were flash-frozen in liquid nitrogen for X-ray diffraction experiments.

Structure determination. X-ray diffraction data sets were collected at the Advanced Photon Source (beamline 24-ID-C), and were processed using the HKL2000 suite of computer programs[51]. The structure of $\Delta P2X_4$-C was initially obtained by molecular replacement with the earlier $\Delta P2X_4$-B_1 coordinates (PDB code 3H9V) using the program Phaser[52]. There is one subunit in the asymmetric unit and the entire trimeric receptor is built-up by subunits related by crystallographic symmetry. The initial model of the extracellular domain was rebuilt and refined using programs in the *CCP4*[53], COOT[54] and PHENIX[55] packages. The higher resolution data set of $\Delta P2X_4$-C confirmed a register shift around the residues 88–97 in the $\Delta P2X_4$-B_1 structure (Supplementary Fig. 5). After iterative cycles of model building and refinement, the electron density map was recalculated using the extracellular domain of the refined $\Delta P2X_4$-C structure. The resulting electron density maps demonstrated that the transmembrane domains associated with the $\Delta P2X_4$-C maps adopted a markedly different conformation from that of the transmembrane domains of the $\Delta P2X_4$-B_1 structure. The $\Delta P2X_4$-C transmembrane domains were rebuilt, and further cycles of model building and refinement were performed. The final model contains residues 36–359.

The new apo structure of the zebrafish $\Delta P2X_4$-B_2 construct was obtained by molecular replacement using the previous $\Delta P2X_4$-B_1 crystal structure[16]. The new model corrected the register shift of residues 88–97 and was iteratively refined to good crystallographic residuals as described above. Although crystals were 'back-soaked' with a Gd^{3+}-free solution and soaked with TNP-ATP, we found no electron density for TNP-ATP in the ATP-binding site. Thus, this structure is that of the apo conformation. The correction of the register shift was also applied to the $\Delta P2X_4$-B_1 coordinates, and after refinement with the same sets of reflections as in the original article[16] the R_{work} and R_{free} values improved with the R_{free} value decreasing from 27.8% to 27.4%. All structures were validated by the computer program PROCHECK[56] and MolProbity[57]. Pore-lining surfaces were calculated using HOLE[58]. The rotation axis in Fig. 6 and its angle were calculated by the Dyndom analysis[59].

Electrophysiology. RNA encoding the crystallized construct of zebrafish *p2x4* (also known as *p2rx4a*), $\Delta P2X_4$-C and the N-terminal EGFP fusion of the $\Delta P2X_4$-C construct were transcribed from pCDNA3.1x plasmids using the mMessage mMachine T7 Ultra kit (Ambion). RNA (2.5–5 ng) was then injected into *Xenopus* oocytes. The recording solution was composed of 100 mM NaCl, 5 mM HEPES, 1 mM $MgCl_2$ and 0.3 mM $CaCl_2$, pH 7.6 (ref. 49). Test solutions containing ATP were freshly prepared each day. Recording electrode pipettes (0.5–2 MΩ) were filled with 3 M KCl. The holding potential was at −80 mV. Currents under two-electrode voltage clamp were recorded using a Axoclamp 2B and 900A, GeneClamp 500 amplifiers (Axon Instruments) and digitized using a Digidata 1440A and pClamp 10 (Molecular Devices). No currents were observed from uninjected oocytes after application of ATP-containing recording solutions.

The dose–response relationship for ATP activation was obtained by measuring peak current amplitudes in response to ATP application. The peak current from the test solution at the reference concentration of ATP (6.5 μM) was measured first and the peak current for each ATP test solution was measured 4 min later and normalized to the peak current evoked by the reference solution. Each ATP concentration was tested on four oocytes. The data were fit to the Hill equation using the Graphpad Prism4 program.

Radioligand-binding experiments. The N-terminal, EGFP-fusion $\Delta P2X_4$-C construct was expressed and purified as described earlier, concentrated and dialysed overnight at 4 °C against three changes of a dialysis buffer (buffer I) containing 20 mM HEPES, pH 7.0, 80 mM NaCl, 20 mM KCl, 15% glycerol and 0.5 mM DDM, and stored at −80 °C before use. Measurements of total ATP binding were obtained by adding EGFP-fusion $\Delta P2X_4$-C to a final concentration of 15 nM in 250 μl dialysis buffer containing 0–293 nM ^{3}H-ATP (Perkin Elmer), in which the hot ATP was diluted with cold ATP in a ratio of 1:4, yielding a final specific activity of 7.5 Ci mmol^{-1}. Samples were incubated at 4 °C overnight and then binding was terminated by filtering through GSWP 02500 nitrocellulose membranes (Millipore) pre-equilibrated with dialysis buffer containing 100 μM cold ATP. The membranes were subsequently washed three times with 2 ml buffer I, transferred to scintillation vials containing 6 ml of Ultima Gold scintillation cocktail (Perkin Elmer) and counted. Estimates of nonspecific binding were obtained by reactions carried out in the presence of 100 μM cold ATP. The determination of specific binding was derived by subtraction of the nonspecific binding from total binding. The entire experiment was performed in triplicate. The data were fit to a rectangular hyperbola using the Graphpad Prism4 program. For the [^{3}H]-ATP competition assay, measurements of total ATP binding were obtained by adding 10 nM of EGFP-fusion $\Delta P2X_4$-C, 10 nM of ^{3}H-ATP (7.5 Ci mmol^{-1}) at the final concentration in 250 μl dialysis buffer containing cold nucleotides. Estimates of nonspecific binding were obtained by reactions carried out in the presence of 100 μM cold ATP. Reactions were terminated by filtering, as described earlier. The entire experiment was performed in triplicate. Data were fit to a sigmoidal dose response equation using the Graphpad Prism4 program.

51. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
52. McCoy, A. J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr. D* **63,** 32–41 (2007).
53. Collaborative Computational Project, Number 4. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
54. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
55. Adams, P. D. *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr. D* **58,** 1948–1954 (2002).
56. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26,** 283–291 (1993).
57. Davis, I. W. *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35,** 375–383 (2007).
58. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Samsom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14,** 354–360 (1996).
59. Hayward, S. & Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50. *J. Mol. Graph. Model.* **21,** 181–183 (2002).

LETTER

doi:10.1038/nature11096

The suppression of star formation by powerful active galactic nuclei

M. J. Page[1], M. Symeonidis[1], J. D. Vieira[2], B. Altieri[3], A. Amblard[4], V. Arumugam[5], H. Aussel[6], T. Babbedge[7], A. Blain[8], J. Bock[2,9], A. Boselli[10], V. Buat[10], N. Castro-Rodríguez[11,12], A. Cava[13], P. Chanial[6], D. L. Clements[7], A. Conley[14], L. Conversi[3], A. Cooray[2,15], C. D. Dowell[2,9], E. N. Dubois[16], J. S. Dunlop[5], E. Dwek[17], S. Dye[18], S. Eales[19], D. Elbaz[6], D. Farrah[16], M. Fox[7], A. Franceschini[20], W. Gear[19], J. Glenn[14,21], M. Griffin[19], M. Halpern[22], E. Hatziminaoglou[23], E. Ibar[24], K. Isaak[25], R. J. Ivison[5,24], G. Lagache[26], L. Levenson[2,9], N. Lu[2,27], S. Madden[6], B. Maffei[28], G. Mainetti[20], L. Marchetti[20], H. T. Nguyen[2,9], B. O'Halloran[7], S. J. Oliver[16], A. Omont[29], P. Panuzzo[6], A. Papageorgiou[19], C. P. Pearson[30,31], I. Pérez-Fournon[11,12], M. Pohlen[19], J. I. Rawlings[1], D. Rigopoulou[30,32], L. Riguccini[6], D. Rizzo[7], G. Rodighiero[20], I. G. Roseboom[5,16], M. Rowan-Robinson[7], M. Sánchez Portal[3], B. Schulz[2,27], D. Scott[22], N. Seymour[1,33], D. L. Shupe[2,27], A. J. Smith[16], J. A. Stevens[34], M. Trichas[35], K. E. Tugwell[1], M. Vaccari[20], I. Valtchanov[3], M. Viero[2], L. Vigroux[29], L. Wang[16], R. Ward[16], G. Wright[24], C. K. Xu[2,27] & M. Zemcov[2,9]

The old, red stars that constitute the bulges of galaxies, and the massive black holes at their centres, are the relics of a period in cosmic history when galaxies formed stars at remarkable rates and active galactic nuclei (AGN) shone brightly as a result of accretion onto black holes. It is widely suspected, but unproved, that the tight correlation between the mass of the black hole and the mass of the stellar bulge[1] results from the AGN quenching the surrounding star formation as it approaches its peak luminosity[2–4]. X-rays trace emission from AGN unambiguously[5], whereas powerful star-forming galaxies are usually dust-obscured and are brightest at infrared and submillimetre wavelengths[6]. Here we report submillimetre and X-ray observations that show that rapid star formation was common in the host galaxies of AGN when the Universe was 2–6 billion years old, but that the most vigorous star formation is not observed around black holes above an X-ray luminosity of 10^{44} ergs per second. This suppression of star formation in the host galaxy of a powerful AGN is a key prediction of models in which the AGN drives an outflow[7–9], expelling the interstellar medium of its host and transforming the galaxy's properties in a brief period of cosmic time.

Measuring star formation in galaxies containing powerful AGN has long been a problem, because the radiation from the AGN outshines that from star formation in almost all wavebands. Of all parts of the electromagnetic spectrum, the far-infrared to millimetre waveband offers the best opportunity to measure star formation in galaxies hosting AGN because, in contrast to strongly star-forming galaxies, AGN emit comparatively little radiation at these wavelengths[10]. The combination of deep X-ray and submillimetre observations therefore offers the best prospects for studying the association of star formation and accretion during the $1 < z < 3$ epoch (2–6 billion years after the Big Bang) when star formation and black hole growth in massive galaxies were at their most vigorous (z is redshift).

The X-ray catalogue of the Chandra Deep Field North (hereafter CDF-N) derives from a series of observations made with the Chandra X-ray observatory with a total of 2×10^6 s exposure time[11]. We restrict the sample to those sources detected in the most penetrating (2–8 keV) band to minimize the influence of obscuration on our results, and we further limit the sample to those sources (64%) for which spectroscopic redshifts are available in the literature[12,13]. Luminosities in the 2–8 keV band were calculated assuming that AGN X-ray spectra are power laws of the form[14] $S_\nu \propto \nu^{-0.9}$ where ν is frequency and S_ν is flux density; the luminosities are not corrected for absorption intrinsic to the AGN or their host galaxies. In order to restrict the X-ray sample to AGN, we have discarded any sources with 2–8 keV luminosity $L_X < 10^{42}$ erg s^{-1}. Submillimetre observations (by the SPIRE[15] instrument on the Herschel Space Observatory) of the CDF-N were carried out in October 2009 as part of the HerMES programme[16]. Maps and source catalogues at wavelengths of 250, 350 and 500 μm were constructed[17]. At the depth of the SPIRE maps, the dominant source of uncertainty in the maps is confusion noise due to the high sky density of sources. For cross-matching with the Chandra source catalogue[11] we chose the 250 μm catalogue, which has the most precise positions, and we used only sources with 250 μm flux densities greater than 18 mJy, which corresponds to a signal-to-noise ratio greater than 3 when the effects of confusion are included[17]. X-ray sources were matched to 250 μm sources within 6 arcsec, corresponding to approximately 95% confidence in the 250 μm positions. The detection statistics are given in Table 1. The expected level of spurious associations between X-ray and 250 μm sources was calculated from the sky density of

[1]Mullard Space Science Laboratory, University College London, Holmbury St Mary, Dorking, Surrey RH5 6NT, UK. [2]California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [3]Herschel Science Centre, European Space Astronomy Centre, Villanueva de la Cañada, 28691 Madrid, Spain. [4]NASA, Ames Research Center, Moffett Field, California 94035, USA. [5]Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [6]Laboratoire AIM-Paris-Saclay, CEA/DSM/Irfu – CNRS – Université Paris Diderot, CE-Saclay, pt courrier 131, F-91191 Gif-sur-Yvette, France. [7]Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK. [8]Department of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK. [9]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA. [10]Laboratoire d'Astrophysique de Marseille, OAMP, Université Aix-Marseille, CNRS, 38 rue Frédéric Joliot-Curie, 13388 Marseille cedex 13, France. [11]Instituto de Astrofísica de Canarias (IAC), E-38200 La Laguna, Tenerife, Spain. [12]Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38205 La Laguna, Tenerife, Spain. [13]Departamento de Astrofísica, Facultad de CC Físicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain. [14]Center for Astrophysics and Space Astronomy 389-UCB, University of Colorado, Boulder, Colorado 80309, USA. [15]Department of Physics and Astronomy, University of California, Irvine, California 92697, USA. [16]Astronomy Centre, Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, UK. [17]Observational Cosmology Laboratory, Code 665, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. [18]School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, UK. [19]School of Physics and Astronomy, Cardiff University, Queens Buildings, The Parade, Cardiff CF24 3AA, UK. [20]Dipartimento di Astronomia, Università di Padova, vicolo Osservatorio, 3, 35122 Padova, Italy. [21]Department of Astrophysical and Planetary Sciences, CASA 389-UCB, University of Colorado, Boulder, Colorado 80309, USA. [22]Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada. [23]ESO, Karl-Schwarzschild-Strasse 2, 85748 Garching bei München, Germany. [24]UK Astronomy Technology Centre, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [25]European Space Research and Technology Centre (ESTEC), Keplerlaan 1, 2201 AZ, Noordwijk, The Netherlands. [26]Institut d'Astrophysique Spatiale (IAS), bâtiment 121, Université Paris-Sud 11 and CNRS (UMR 8617), 91405 Orsay, France. [27]Infrared Processing and Analysis Center, MS 100-22, California Institute of Technology, JPL, Pasadena, California 91125, USA. [28]School of Physics and Astronomy, The University of Manchester, Alan Turing Building, Oxford Road, Manchester M13 9PL, UK. [29]Institut d'Astrophysique de Paris, UMR 7095, CNRS, UPMC Université Paris 06, 98 bis boulevard Arago, F-75014 Paris, France. [30]RAL Space, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, UK. [31]Institute for Space Imaging Science, University of Lethbridge, Lethbridge, Alberta T1K 3M4, Canada. [32]Department of Astrophysics, Denys Wilkinson Building, University of Oxford, Keble Road, Oxford OX1 3RH, UK. [33]CSIRO Astronomy and Space Science, PO Box 76, Epping, New South Wales 1710, Australia. [34]Centre for Astrophysics Research, University of Hertfordshire, College Lane, Hatfield, Hertfordshire AL10 9AB, UK. [35]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

Table 1 | 250 μm detection statistics in various regions of parameter space

Region of (z,L_X) parameter space	Number of AGN	Number of AGN associated with 250 μm sources	Expected number of spurious associations	Fraction of AGN associated with 250 μm sources
All z, $10^{42}\,\mathrm{erg\,s^{-1}} < L_X < 10^{45}\,\mathrm{erg\,s^{-1}}$	176	24	2.1	$14 \pm 3\,(^{+6}_{-5})\%$
$1 < z < 3$, $10^{43}\,\mathrm{erg\,s^{-1}} < L_X < 10^{44}\,\mathrm{erg\,s^{-1}}$	44	11	0.5	$25^{+8}_{-7}\,(^{+15}_{-12})\%$
$1 < z < 3$, $10^{44}\,\mathrm{erg\,s^{-1}} < L_X < 10^{45}\,\mathrm{erg\,s^{-1}}$	21	0	0.2	$<5(<13)\%$

The first row corresponds to the entire sample of secure AGN in the CDF-N, while the second and third rows correspond to the regions enclosed within the blue dashed lines in Fig. 1. Confidence intervals on the fraction of AGN associated with 250 μm sources are given at 68%, with 95% intervals enclosed in brackets. It should be noted that there is one case of two AGN being associated with the same 250 μm source. The two AGN have very similar spectroscopic redshifts, and both have X-ray luminosities between 10^{43} and $10^{44}\,\mathrm{erg\,s^{-1}}$. Although the two AGN cannot be resolved at 250 μm, source extraction using X-ray and 24 μm positions as priors[25] indicates that both AGN are 5σ sources at 250 μm.

250 μm sources in annular regions of radius 10–30 arcsec around the X-ray source positions, and is reported in Table 1.

The distribution of CDF-N AGN in the redshift–X-ray luminosity (z–L_X) plane is shown in Fig. 1, and reveals a striking trend of 250 μm detectability with X-ray luminosity: of the 24 AGN detected at 250 μm, none of them have $L_X > 10^{44}\,\mathrm{erg\,s^{-1}}$. The redshift range between 1 and 3 is of most interest, because it corresponds to the epoch in which powerful AGN accreted most of their black hole mass and present-day massive galaxies formed most of their stars. Within this redshift range, Fig. 1 shows that 11 out of 44 AGN ($25^{+8}_{-7}\%$) with $10^{43}\,\mathrm{erg\,s^{-1}} < L_X < 10^{44}\,\mathrm{erg\,s^{-1}}$ are detected at 250 μm, while none of the 21 objects with $L_X > 10^{44}\,\mathrm{erg\,s^{-1}}$ are detected. The difference

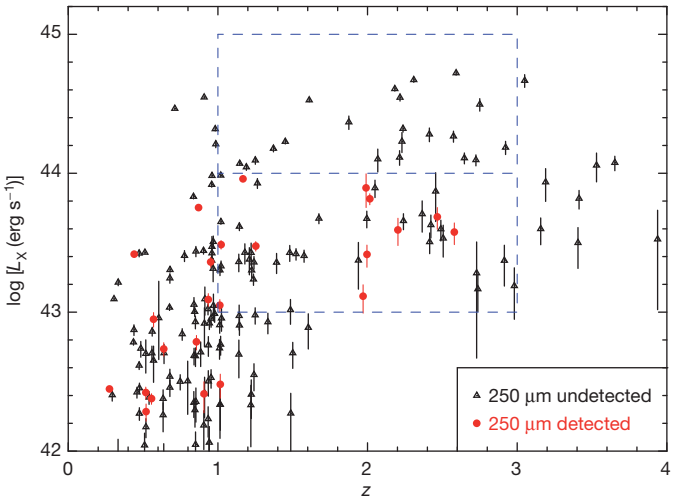


Figure 1 | Redshifts (z) and 2–8 keV X-ray luminosities (L_X) of AGN in the CDF-N. The luminosities have been corrected to the rest frame assuming a spectrum $S_\nu \propto \nu^{-0.9}$ and are not corrected for intrinsic absorption. The blue dashed rectangles delimit the luminosity decades above and below $10^{44}\,\mathrm{erg\,s^{-1}}$ in the $1 < z < 3$ redshift range. Error bars, 68% confidence limits.

in detection rates has a significance of >99%, according to a single-tail Fisher's exact test. We have considered the effects that incompleteness in the spectroscopic redshifts, or absorption of the X-ray flux by gas and dust, might have on our results. We find that the systematic non-detection of the powerful AGN is robust against both effects, although X-ray absorption does appear to be a common property of the AGN detected at 250 μm. We have also verified the low 250 μm detection rate of AGN with $L_X > 10^{44}\,\mathrm{erg\,s^{-1}}$ using the Extended Chandra Deep Field South field, finding that of 49 such sources with $1 < z < 3$, only 1 is detected at 250 μm.

Infrared spectral energy distributions for the 250-μm-detected AGN were constructed by combining the SPIRE photometry with 3.6–160 μm photometry from the Spitzer Space Telescope. X-ray and infrared properties of the 11 250-μm-detected AGN with $1 < z < 3$ and L_X in the range 10^{43}–$10^{44}\,\mathrm{erg\,s^{-1}}$ are given in Table 2. In most cases, the AGN contributes less than 10% to the infrared luminosity. The best-fit infrared luminosities lie between 4×10^{11} and 10^{13} times solar luminosity ($L_\odot$), implying star formation rates between 50 and 1,750 solar masses ($M_\odot$) per year[18].

We performed a stacking analysis for the $1 < z < 3$ AGN to probe below the confusion limit of the SPIRE images. We split the sample into five bins of L_X from 10^{43} to $10^{45}\,\mathrm{erg\,s^{-1}}$ and determined the average star formation rates of AGN in each bin. The results are shown in Fig. 2. In the redshift range $1 < z < 3$, the mean star formation rate in AGN with L_X of 10^{43}–$10^{44}\,\mathrm{erg\,s^{-1}}$ is $214 \pm 25\ M_\odot$ per year, compared to a mean star formation rate for AGN with $L_X > 10^{44}\,\mathrm{erg\,s^{-1}}$ of $65 \pm 18\ M_\odot$ per year. These averages are independent of the SPIRE 250 μm detection limit because they are obtained from a stack of all sources within a given range of L_X, whether detected at 250 μm or not.

At redshifts of 1–3, the X-ray luminosity of $10^{44}\,\mathrm{erg\,s^{-1}}$, which divides the regions of 250 μm detection and non-detection in Fig. 1, corresponds approximately to the knee in the luminosity function of AGN[19]. The steep shape of the luminosity function at $L_X > 10^{44}\,\mathrm{erg\,s^{-1}}$ implies that this part of the luminosity function is dominated by objects which are at the peak of their accretion rates. Our observations

Table 2 | Properties of SPIRE-detected AGN

ID	Redshift	Log[L_X (erg s^{-1})]	Log[N_H (atoms cm^{-2})]	Absorption correction	Log[L_{IR} ($L_\odot$)]	AGN (%)	SFR ($M_\odot$ yr^{-1})
35	2.203	$43.59^{+0.08}_{-0.11}$	$23.6^{+0.1}_{-0.2}$	$0.15^{+0.05}_{-0.07}$	12.70 ± 0.03	12	750–850
109	2.580	$43.58^{+0.07}_{-0.09}$	$23.4^{+0.1}_{-\mathrm{unc}}$	$0.09^{+0.03}_{-0.09}$	13.01 ± 0.05	5	1,660–1,750
135	2.466	$43.69^{+0.07}_{-0.09}$	>24.0	>0.30	12.81 ± 0.05	4	1,060–1,110
158	1.013	$43.05^{+0.04}_{-0.05}$	$23.01^{+0.1}_{-0.1}$	$0.15^{+0.02}_{-0.02}$	12.29 ± 0.09	4	320–330
190	2.015	$43.81^{+0.04}_{-0.04}$	$23.6^{+0.1}_{-0.1}$	$0.16^{+0.02}_{-0.02}$	12.88 ± 0.03	21	1,030–1,300
331	1.253	$43.48^{+0.03}_{-0.04}$			12.51 ± 0.07	5	530–550
366	1.970	$43.11^{+0.08}_{-0.12}$	$23.4^{+0.1}_{-0.2}$	$0.11^{+0.03}_{-0.04}$	12.84 ± 0.05	3	1,140–1,170
368	1.996	$43.42^{+0.07}_{-0.09}$	>23.8	>0.26	12.41 ± 0.06	3	420–430
384	1.021	$43.49^{+0.03}_{-0.03}$	$23.4^{+0.1}_{-0.1}$	$0.25^{+0.02}_{-0.03}$	11.55 ± 0.17	11	50–60
455	1.168	$43.96^{+0.02}_{-0.02}$			11.97 ± 0.04	30	110–160
500	1.990	$43.89^{+0.10}_{-0.14}$	$23.2^{+0.2}_{-\mathrm{unc}}$	$0.07^{+0.03}_{-0.07}$	12.62 ± 0.03	4	690–710

Data are given for AGN with $1 < z < 3$ and $10^{43}\,\mathrm{erg\,s^{-1}} < L_X < 10^{44}\,\mathrm{erg\,s^{-1}}$. The first column gives the ID number of the source in the X-ray catalogue[11]. The second column gives the redshift, and the third column gives the logarithm of the 2–8 keV X-ray luminosity (L_X). The fourth column gives the logarithm of the column density of absorbing gas (N_H, in units of hydrogen atoms per cm^2) implied by the ratio of 2–8 keV to 0.5–2 keV X-rays; a blank entry indicates no evidence for photoelectric absorption in X-rays, and 'unc' is used where the lower limit to the column density is unconstrained. The fifth column gives the correction to log L_X to account for the absorption. The sixth column gives the 8–1,000 μm infrared luminosity, L_{IR}. The seventh column gives the maximum likely contribution of an AGN to the infrared luminosity, and the eighth column gives the range of star formation rate (SFR) implied by the infrared luminosity, where the upper and lower limits correspond to zero AGN contribution and the maximum AGN contribution to the infrared luminosity, respectively. Photometry for the spectral energy distributions was extracted from Spitzer and SPIRE images using the X-ray and 24 μm catalogue positions as priors[25]. Total 8–1,000 μm infrared luminosities were then determined by fitting templates[26] to the spectral energy distributions[27]. Upper limits to the AGN contribution to the infrared luminosities were obtained by normalizing an AGN template in the mid-infrared[28].

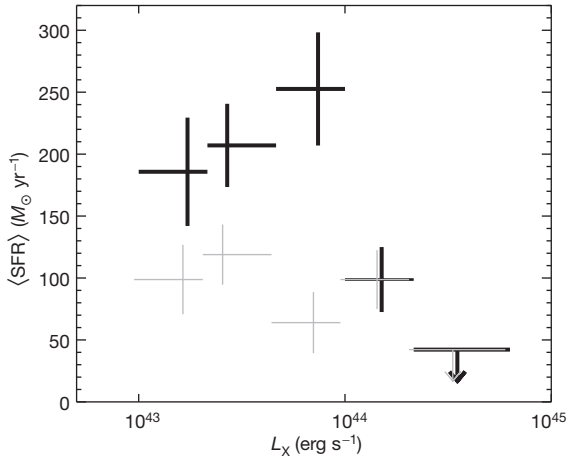


Figure 2 | Average star formation rates, ⟨SFR⟩, derived from averaged far-infrared luminosities of $1 < z < 3$ AGN, as a function of L_X. We converted the 250, 350 and 500 μm flux densities for each source into an equivalent 8–1,000 μm luminosity by fitting a grey-body curve, with a temperature of 30 K in the rest-frame of the source, an emissivity index of $\beta = 1.6$, and a power-law extension to the Wien side[29] and multiplying by $4\pi D_L^2$, where D_L is the luminosity distance. Fluctuations in the map sometimes scatter the fluxes of undetected sources to negative values, which translate to negative luminosities when multiplied by $4\pi D_L^2$. Such negative solutions for individual AGN were allowed so as not to produce an artificial positive bias in the averages. The luminosities were averaged in five bins in L_X, which were chosen to include a similar number of AGN in each bin. The average luminosities were then converted to star formation rates[18]. AGN which are individually detected at 250 μm are included in the averages shown in bold black, but have been excluded from the averages which are shown in grey, to show the contribution that these sources make to the average star formation rates. The grey points have been offset horizontally from the bold black points for clarity. Error bars correspond to 68% confidence limits and were determined by bootstrap resampling, with a 7% systematic error added in quadrature to account for the calibration error on SPIRE photometry.

therefore imply that the most prodigious episodes of star formation are common in the host galaxies of $1 < z < 3$ AGN, but avoid powerful AGN in which accretion is at its peak.

This systematic non-coincidence of the peak periods of star formation and accretion implies a direct interaction between the two processes, and provides a powerful discriminator for the form of AGN feedback which is responsible for terminating star formation in the host galaxy. Two families of feedback models have been proposed, widely referred to as 'quasar mode' and 'radio mode'[20]. In quasar-mode feedback, a luminous AGN generates a powerful wind which terminates star formation by driving the interstellar medium from the surrounding host galaxy. In radio-mode feedback, star formation is suppressed because collimated jets of relativistic particles emitted by a radiatively inefficient AGN prevent gas in the surrounding hot halo from cooling, thereby starving the galaxy of cool gas from which to form stars.

Radio-mode feedback is commonly invoked in semi-analytical models to limit galaxy masses and luminosities[20,21]. In these models, black holes grow through luminous accretion episodes and black hole mergers. The correlation between black hole mass and bulge mass comes from assuming that a fixed fraction of the gas is accreted by the nucleus during each star forming episode that results from a galaxy merger or disc instability, and hence star formation and accretion rate should be correlated over the full range of luminosity. Our observations are therefore inconsistent with models in which AGN influence their host galaxies only through radio-mode feedback[20,21]. In contrast, models of galaxy formation in which quasar-mode feedback is responsible for terminating the star formation[7,8,9,22], and which have received some observational support recently[23,24], predict that the AGN luminosity peaks later than the star formation rate, and thus are consistent with our observations. These models also predict that residual star formation, at the level of a few tens of per cent of the peak, will continue during the period in which the AGN luminosity is at its maximum, consistent with our stacked results; our results show that, on average, AGN with $L_X > 10^{44}$ erg s^{-1} are still forming stars at approximately 65 $M_\odot$ per year. Our observations do not discriminate between models invoking major mergers[8] or accretion of gas into a massive halo[22] as the trigger for the intense star formation. After the interstellar medium has been driven out by the luminous AGN and the AGN itself becomes starved of fuel, radio-mode feedback is the most credible agent by which further star formation is inhibited.

Received 30 November 2011; accepted 29 March 2012.

1. Häring, N. & Rix, H. W. On the black hole mass-bulge mass relation. *Astrophys. J.* **604,** L89–L92 (2004).
2. Silk, J. & Rees, M. J. Quasars and galaxy formation. *Astron. Astrophys.* **331,** L1–L4 (1998).
3. Fabian, A. C. The obscured growth of massive black holes. *Mon. Not. R. Astron. Soc.* **308,** L39–L43 (1999).
4. King, A. R. Black hole outflows. *Mon. Not. R. Astron. Soc.* **402,** 1516–1522 (2010).
5. Brandt, W. N. & Hasinger, G. Deep extragalactic X-ray surveys. *Annu. Rev. Astron. Astrophys.* **43,** 827–859 (2005).
6. Sanders, D. B. & Mirabel, I. F. Luminous infrared galaxies. *Annu. Rev. Astron. Astrophys.* **34,** 749–792 (1996).
7. Di Matteo, T., Springel, V. & Hernquist, L. Energy input from quasars regulates the growth and activity of black holes and their host galaxies. *Nature* **433,** 604–607 (2005).
8. Springel, V., Di Matteo, T. & Hernquist, L. Modelling feedback from stars and black holes in galaxy mergers. *Mon. Not. R. Astron. Soc.* **361,** 776–794 (2005).
9. Sijacki, D., Springel, V., Di Matteo, T. & Herhquist, L. A unified model for AGN feedback in cosmological simulations of structure formation. *Mon. Not. R. Astron. Soc.* **380,** 877–900 (2007).
10. Hatziminaoglou, E. *et al.* HerMES: far infrared properties of known AGN in the HerMES fields. *Astron. Astrophys.* **518,** L33 (2010).
11. Alexander, D. M. *et al.* The Chandra Deep Field North Survey. XIII. 2 Ms point-source catalogs. *Astron. J.* **126,** 539–574 (2003).
12. Trouille, L., Barger, A. J., Cowie, L. L., Yang, Y. & Mushotzky, R. F. The OPTX Project. I. The flux and redshift catalogs for the CLANS, CLASXS, and CDF-N fields. *Astrophys. J.* **179** (Suppl.), 1–18 (2008).
13. Barger, A. J. *et al.* A highly complete spectroscopic survey of the GOODS-N field. *Astrophys. J.* **689,** 687–708 (2008).
14. Mateos, S. *et al.* XMM-Newton observations of the Lockman Hole IV: spectra of the brightest AGN. *Astron. Astrophys.* **444,** 79–99 (2005).
15. Griffin, M. J. *et al.* The Herschel-SPIRE instrument and its in-flight performance. *Astron. Astrophys.* **518,** L3 (2010).
16. Oliver, S. J. *et al.* HerMES: SPIRE galaxy number counts at 250, 350, and 500 μm. *Astron. Astrophys.* **518,** L21 (2010).
17. Smith, A. J. *et al.* HerMES: point source catalogues from deep Herschel-SPIRE observations. *Mon. Not. R. Astron. Soc.* **419,** 377–389 (2012).
18. Kennicutt, R. C. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498,** 541–552 (1998).
19. Ebrero, J. *et al.* The XMM-Newton serendipitous survey. VI. The X-ray luminosity function. *Astron. Astrophys.* **493,** 55–69 (2009).
20. Croton, D. J. *et al.* The many lives of active galactic nuclei: cooling flows, black holes and the luminosities and colours of galaxies. *Mon. Not. R. Astron. Soc.* **365,** 11–28 (2006).
21. Bower, R. G. *et al.* Breaking the hierarchy of galaxy formation. *Mon. Not. R. Astron. Soc.* **370,** 645–655 (2006).
22. Granato, G. L. *et al.* A physical model for the coevolution of QSOs and their spheroidal hosts. *Astrophys. J.* **600,** 580–594 (2004).
23. Farrah, D. *et al.* Direct evidence for termination of obscured star formation by radiatively driven outflows in reddened QSOs. *Astrophys. J.* **745,** 178 (2012).
24. Cano-Díaz, M. *et al.* Observational evidence of quasar feedback quenching star formation at high redshift. *Astron. Astrophys.* **537,** L8 (2012).
25. Roseboom, I. G. *et al.* The Herschel Multi-Tiered Extragalactic Survey: source extraction and cross-identifications in confusion-dominated SPIRE images. *Mon. Not. R. Astron. Soc.* **409,** 48–65 (2010).
26. Siebenmorgen, R. & Krugel, E. Dust in starburst nuclei and ULIRGs. SED models for observers. *Astron. Astrophys.* **461,** 445–453 (2007).
27. Symeonidis, M. *et al.* The link between SCUBA and Spitzer: cold galaxies at $z \leq 1$. *Mon. Not. R. Astron. Soc.* **397,** 1728–1738 (2009).
28. Seymour, N. *et al.* HerMES: SPIRE emission from radio-selected active galactic nuclei. *Mon. Not. R. Astron. Soc.* **413,** 1777–1786 (2011).
29. Blain, A. W., Barnard, V. E. & Chapman, S. C. Submillimetre and far-infrared spectral energy distributions of galaxies: the luminosity-temperature relation and consequences for photometric redshifts. *Mon. Not. R. Astron. Soc.* **338,** 733–744 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Herschel is an ESA space observatory with science instruments provided by European-led Principal Investigator consortia and with important participation from NASA. SPIRE has been developed by a consortium of institutes led

by Cardiff University (UK) and which includes: University of Lethbridge (Canada); NAOC (China); CEA, LAM (France); IFSI, University of Padua (Italy); IAC (Spain); Stockholm Observatory (Sweden); Imperial College London, RAL, UCL-MSSL, UKATC, University of Sussex (UK); and Caltech, JPL, NHSC, University of Colorado (USA). This development has been supported by national funding agencies: CSA (Canada); NAOC (China); CEA, CNES, CNRS (France); ASI (Italy); MCINN (Spain); SNSB (Sweden); STFC, UKSA (UK); and NASA (USA).

Author Contributions This Letter represents the combined work of the HerMES collaboration, the SPIRE Instrument Team's extragalactic survey. M.J.P. planned the study, and wrote the draft version of the paper. M.S. fitted models to the spectral energy distributions of the sources and J.D.V. performed the stacking analysis. All other co-authors contributed extensively and equally by their varied contributions to the SPIRE instrument, the Herschel mission, analysis of SPIRE and HerMES data, planning of HerMES observations and scientific support of HerMES, and by commenting on this manuscript as part of an internal review process.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.P. (mjp@mssl.ucl.ac.uk).

LETTER

doi:10.1038/nature10990

An ultraviolet–optical flare from the tidal disruption of a helium-rich stellar core

S. Gezari[1], R. Chornock[2], A. Rest[3], M. E. Huber[4], K. Forster[5], E. Berger[2], P. J. Challis[2], J. D. Neill[5], D. C. Martin[5], T. Heckman[1], A. Lawrence[6], C. Norman[1], G. Narayan[2], R. J. Foley[2], G. H. Marion[2], D. Scolnic[1], L. Chomiuk[2], A. Soderberg[2], K. Smith[7], R. P. Kirshner[2], A. G. Riess[1], S. J. Smartt[7], C. W. Stubbs[2], J. L. Tonry[4], W. M. Wood-Vasey[8], W. S. Burgett[4], K. C. Chambers[4], T. Grav[9], J. N. Heasley[4], N. Kaiser[4], R.-P. Kudritzki[4], E. A. Magnier[4], J. S. Morgan[4] & P. A. Price[10]

The flare of radiation from the tidal disruption and accretion of a star can be used as a marker for supermassive black holes that otherwise lie dormant and undetected in the centres of distant galaxies[1]. Previous candidate flares[2–6] have had declining light curves in good agreement with expectations, but with poor constraints on the time of disruption and the type of star disrupted, because the rising emission was not observed. Recently, two 'relativistic' candidate tidal disruption events were discovered, each of whose extreme X-ray luminosity and synchrotron radio emission were interpreted as the onset of emission from a relativistic jet[7–10]. Here we report a luminous ultraviolet–optical flare from the nuclear region of an inactive galaxy at a redshift of 0.1696. The observed continuum is cooler than expected for a simple accreting debris disk, but the well-sampled rise and decay of the light curve follow the predicted mass accretion rate and can be modelled to determine the time of disruption to an accuracy of two days. The black hole has a mass of about two million solar masses, modulo a factor dependent on the mass and radius of the star disrupted. On the basis of the spectroscopic signature of ionized helium from the unbound debris, we determine that the disrupted star was a helium-rich stellar core.

When the pericentre of a star's orbit (R_p) passes within the tidal disruption radius of a massive black hole, $R_T \approx R_*(M_{BH}/M_*)^{1/3}$ (where R_* is the stellar radius, M_{BH} is the black-hole mass and M_* is the stellar mass), tidal forces overcome the binding energy of the star, which breaks up with roughly half of the stellar debris remaining bound to the black hole and the rest being ejected at high velocity[1]. For black holes above a critical mass, $M_{crit} \approx 10^8 r_*^{3/2} m_*^{-1/2} M_\odot$ (where $r_* = R_*/R_\odot$, $m_* = M_*/M_\odot$, $R_\odot$ is the solar radius and $M_\odot$ is the solar mass), the star becomes trapped within the event horizon of the black hole before being disrupted. The mass accretion rate ($\dot{M}$) in a tidal disruption event can be calculated directly from the orbital return times of the bound debris[1,11,12]. For the simplest case, of a star of uniform density, this yields $\dot{M} = (2/3)(fM_*/t_{min})(t/t_{min})^{-5/3}$, where f is the fraction of the star accreted and t_{min} is the orbital period of the most tightly bound debris and, therefore, the time delay between the time of disruption and the start of the flare, which scales as $M_{BH}^{1/2} M_*^{-1} R_*^{3/2}$ for $R_p = R_T$. The radiative output of the accreted debris is less certain, and depends on the ratio of the accretion rate to the Eddington rate[13].

The optical transient, PS1-10jh (right ascension, $\alpha_{J2000} = 16\,h\ 09\,min\ 28.296\,s$; declination, $\delta_{J2000} = +53^\circ\ 40'\ 23.52''$), was discovered on 2010 May 31.45 UT (universal time) in the Pan-STARRS1[14] (PS1) Medium Deep Survey by our two independent image-differencing pipelines. The densely sampled (cadence, $\Delta t \approx 3$ d) optical light curves of PS1-10jh in the g_{P1}, r_{P1}, i_{P1} and z_{P1} bands (Supplementary Information) follow the rise of the transient to its peak in the g_{P1} band on 2010 July 12.31 UT and its subsequent decay until 2011 September 1.24 UT (Supplementary Table 1). PS1-10jh was discovered independently as a transient, near-ultraviolet (NUV) source at the 20σ level by the Galaxy Evolution Explorer[15] (GALEX) Time Domain Survey (TDS) on 2010 June 17.68 UT within 2.5 ± 3.0 arcsec of the PS1 location, and was detected in ten more epochs of TDS observations between then and 2011 June 10.68 UT (Supplementary Table 2). No source is detected in a coaddition of all the TDS epochs in 2009, with a 3σ upper limit of >25.6 mag implying a peak amplitude of variability in the NUV of >6.4 mag. See Supplementary Information for details on the PS1 and GALEX photometry.

PS1-10jh is coincident with the centre of a galaxy within the 3σ positional uncertainty (0.036 arcsec; Supplementary Information), with rest-frame u-, g-, r-, i- and z-band photometry from the Sloan Digital Sky Survey[16] and K-band photometry from the UK Infrared Telescope Infrared Sky Survey[17] fitted with a galaxy template[18] with $M_{stars} = (3.6 \pm 0.2) \times 10^9 M_\odot$ and $M_r = -18.7$ mag, where M_{stars} is the galaxy stellar mass and M_r is the absolute r-band magnitude. The mass of the central black hole as determined indirectly from locally established scaling relations[19] is $4^{+4}_{-2} \times 10^6 M_\odot$. We obtained five epochs of optical spectroscopy at the location of PS1-10jh between 2010 June 16.33 and 2011 September 4.23 UT with the 6.5-m MMT (Supplementary Table 3). The continua in the spectra are well modelled by the combination of a galaxy host at redshift $z = 0.1696$ (luminosity distance, 816 Mpc) with a stellar population with an age of 1.4–5.0 Gyr, depending on the chosen metallicity, and a fading hot blackbody component with temperature $T_{BB} \approx 3 \times 10^4$ K (Fig. 1).

The spectra show no narrow emission lines that would be indicative of star formation or an active galactic nucleus (AGN). We obtained a 10-ks, 0.2–10-keV X-ray observation, using the Chandra X-ray Observatory, at the location of PS1-10jh on 2011 May 22.96 UT, and detected no source above the background with a 3σ upper limit of $L_X(0.2\text{–}10\,keV) < 5.8 \times 10^{41}\,erg\,s^{-1}$ for an unobscured AGN spectrum. The X-ray faintness and extreme NUV variability amplitude of PS1-10jh strongly disfavour its origin in an AGN, and its prolonged brightness in the ultraviolet strongly disfavours its origin in a supernova (Supplementary Information).

The rise and decay of the light curve of PS1-10jh is well described by numerical simulations of the mass return rate from a star that is tidally disrupted at $R_p = R_T$ and has an internal structure parameterized by a polytropic exponent of 5/3 characteristic of a fully convective star or a degenerate core[20] (Fig. 2). The decay from the peak is too steep to be fitted by simulations of a more centrally concentrated stellar structure, such as one that is characteristic of a solar-type star (Supplementary Information). There are systematic differences between the light curve

[1]Department of Physics and Astronomy, Johns Hopkins University, 3400 North Charles Street, Baltimore, Maryland 21218, USA. [2]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA. [3]Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA. [4]Institute for Astronomy, University of Hawaii, 2680 Woodlawn Drive, Honolulu, Hawaii 96822, USA. [5]California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA. [6]Institute for Astronomy, University of Edinburgh Scottish Universities Physics Alliance, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK. [7]Astrophysics Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, UK. [8]Pittsburgh Particle Physics, Astrophysics, and Cosmology Center, Department of Physics and Astronomy, University of Pittsburgh, 3941 O'Hara Street, Pittsburgh, Pennsylvania 15260, USA. [9]Planetary Science Institute, 1700 East Fort Lowell, Tucson, Arizona 85719, USA. [10]Department of Astrophysical Sciences, Princeton University, Princeton, New Jersey 08544, USA.

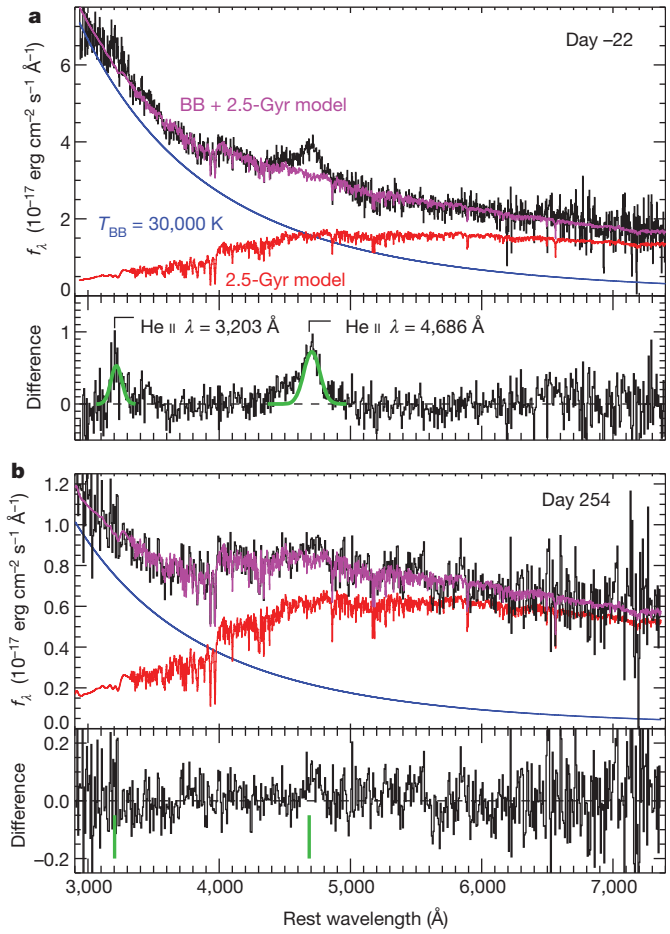


Figure 1 | Optical spectrum. MMT optical spectra (black) of PS1-10jh obtained −22 (**a**) and 254 (**b**) rest-frame days from the peak, expressed in terms of flux density. Each continuum is fitted with a combination (magenta) of a stellar population 2.5 Gyr old (red) and a fading blackbody with a temperature of $\sim 3 \times 10^4$ K determined from the ultraviolet–optical spectral energy distribution (SED) (blue). The difference between the black and magenta spectra is shown in the lower part of each panel. Helium II emission at $\lambda = 4{,}686$ Å (Fowler series, $n = 4 \rightarrow 3$) is detected above the continuum model and fitted with a Gaussian with a full-width at half-maximum of $9{,}000 \pm 700$ km s^{-1} and luminosity $L = (9 \pm 1) \times 10^{40}$ erg s^{-1} (plotted with a green line in the early epoch (**a**)). Residual emission above the continuum model is also detected at ~3,200 Å, which is coincident with the location of the He II $\lambda = 3{,}203$ Å (Fowler series, $n = 5 \rightarrow 3$) line, and confirms the identification of He II $\lambda = 4{,}686$ Å emission. The observed flux ratio of He II $\lambda = 3{,}203$ Å emission to He II $\lambda = 4{,}686$ Å emission is 0.50 ± 0.10, measured using a Gaussian fit to the $\lambda = 3{,}203$ Å line with a width fixed to that of the $\lambda = 4{,}686$ Å line, limits the internal extinction to $E(B-V) < 0.08$ mag (Supplementary Information). The He II $\lambda = 4{,}686$ Å line is still evident as an excess above the model in the later epoch (**b**), but it has faded by a factor of ~10 since 22 rest-frame days before the peak, the same factor by which the ultraviolet continuum has faded during this time. The absolute flux scaling in the later epoch is uncertain owing to obscuration by clouds on the date of the observation.

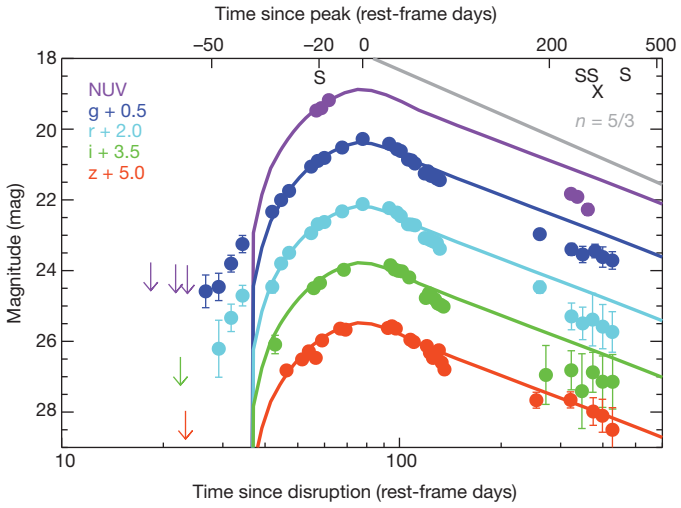


Figure 2 | Ultraviolet–optical light curve. The GALEX NUV and PS1 g_{P1}-, r_{P1}-, i_{P1}- and z_{P1}-band light curves of PS1-10jh (with the flux from the host galaxy removed), plotted against logarithmic time since the peak (top axis) and since the disruption (bottom axis).The curves (shown with solid lines scaled to the flux in the GALEX and PS1 bands) were determined from the best fit of the g_{P1}-band light curve to a numerical model[20] of the mass accretion rate of a tidally disrupted star with a polytropic exponent of 5/3. For each of the four optical bands, we independently performed a least-squares fit of the model for a $10^6 M_\odot$ black hole to the light curve from −36 to 58 rest-frame days from the peak, with the time of disruption, a vertical scaling factor and a time stretch factor as free parameters. The GALEX and PS1 photometry at $t > 240$ rest-frame days since the peak is shown binned in time to increase the signal-to-noise ratio. The dates of multiple epochs of MMT spectroscopy are marked with an S, and the date of the Chandra X-ray observation is marked with an X. The grey line shows an $n = 5/3$ power-law decay from the peak. Errors, 1σ; arrows, 3σ upper limits.

and the model during the early rise (more than 44 rest-frame days before the peak) and the late decay (more than 240 rest-frame days after the peak), which could imply a stellar structure more complex than one described by a single polytrope. The mass of the black hole is determined from the stretch factor of 1.38 ± 0.03 applied to fit the model of a $10^6 M_\odot$ black hole to the light curve, which implies that the time of disruption was 76 ± 2 d before the peak and that $M_{BH} = (1.9 \pm 0.1) \times 10^6 m_*^2 r_*^{-3} M_\odot$.

The most constraining property of PS1-10jh is the detection of very broad high-ionization He II emission at wavelengths of $\lambda = 4{,}686$ Å (full-width at half-maximum, $9{,}000 \pm 700$ km s^{-1}) and 3,203 Å that fade in time along with the ultraviolet–optical continuum. The lack of Balmer line emission in the spectra requires an extremely low hydrogen mass fraction, of <0.2 (Supplementary Information), which cannot be found in the ambient interstellar medium or in a passive accretion disk. This is the strongest evidence that PS1-10jh must be fuelled by the accretion of a star that has lost its hydrogen envelope, either through stellar winds or through tidal interactions with the central supermassive black hole. The broad width of the line is also what is expected from the velocities of the most energetic unbound stellar debris in a tidal disruption event[21], that is $v_{max} \approx 1 \times 10^4 (M_{BH}/10^6 M_\odot)^{1/6} (R_T/R_p) r_*^{-1/2} m_*^{1/3}$ km s^{-1}

We measure the SED of the flare over time from the nearly simultaneous PS1 optical and GALEX ultraviolet observations (with the host galaxy flux removed; Fig. 3). The pre-peak SED is fitted by a blackbody with $T_{BB} = (2.9 \pm 0.2) \times 10^4$ K, consistent with the blackbody component seen in the spectra. However, the temperature fit is very sensitive to internal extinction. If we correct for the maximum internal extinction of $E(B-V) = 0.08$ mag allowed by the ratio between the observed He II $\lambda = 3{,}203$ Å and $\lambda = 4{,}686$ Å emission, the best-fit temperature increases to $(5.5 \pm 0.4) \times 10^4$ K. In fact, we know that the photo-ionizing continuum must have $T_{BB} \gtrsim 5 \times 10^4$ K 22 rest-frame days before the peak to produce enough $\lambda < 228$ Å photons to photo-ionize the He II $\lambda = 4{,}686$ Å line observed with a luminosity of $(9 \pm 1) \times 10^{40}$ erg s^{-1}. The late-time SED can be fitted with the same temperature as the pre-peak SED. We note that the observed continuum temperature, and even the maximum temperature allowed by possible de-reddening, is considerably cooler than the temperature of $\sim 2.5 \times 10^5 (M_{BH}/10^6 M_\odot)^{1/12} r_*^{-1/2} m_*^{-1/6}$ K expected from material radiating at the Eddington limit at the tidal radius[13]. This discrepancy is also seen in AGNs[22] and may imply that the continuum we see is due to reprocessing of some kind[22,23].

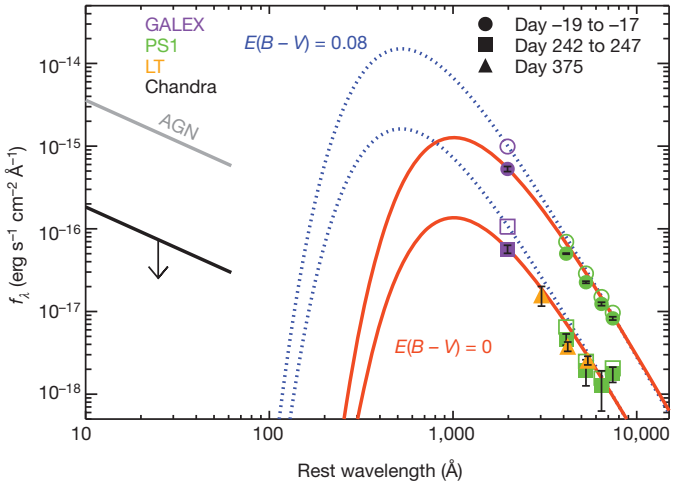


Figure 3 | Spectral energy distribution. SED of PS1-10jh during nearly simultaneous GALEX ultraviolet and PS1 optical observations (with the flux from the host galaxy removed) at two epochs (rest-frame days −19 to −17 and 242 to 247 from the peak of the flare). Flux densities have been corrected for galactic extinction of $E(B-V)=0.013$ mag. The ultraviolet–optical SED from −19 to 247 rest-frame days from the peak is fitted with a 2.9×10^4 K blackbody. Orange solid lines show blackbodies with this temperature scaled to the respective NUV flux densities for the respective epochs. Open symbols show the GALEX and PS1 flux densities corrected for an internal extinction of $E(B-V)=0.08$ mag, and the dotted blue lines shows the 5.5×10^4 K blackbody fit to the de-reddened flux densities for the respective epochs. The upper limit from the Chandra observation on 2011 May 22.96 UT assuming a spectrum with a photon index of $\Gamma=2$, typical of an AGN, is plotted with a thick black line. The X-ray flux density expected from an unobscured AGN with a comparable NUV flux is plotted for comparison with a thick grey line[28]. Also shown are the u-, g- and r-band flux densities measured from aperture photometry with the Liverpool Telescope[29] (LT) on 2011 September 24.91 UT, after subtracting the flux from the host galaxy as measured by the Sloan Digital Sky Survey. Errors, 1σ.

On the basis of the arguments above, we assume that the observed temperature is a lower limit, $T_{\rm BB}\gtrsim3\times10^4$ K. The peak bolometric luminosity is thus $\gtrsim2.2\times10^{44}$ erg s^{-1} and the total energy emitted from integrating under the light-curve model is $\gtrsim2.1\times10^{51}$ erg, corresponding to a total accreted mass ($M_{\rm acc}$) of $\gtrsim0.012(\varepsilon/0.1)^{-1}\,M_\odot$, where ε is the efficiency of converting matter into radiation.

The internal structure and high helium abundance of the star derived from the light curve and the spectra can be consistently modelled as the tidally stripped core of a red giant (the precursor to a helium white dwarf) that had a main-sequence mass of $M_*\gtrsim1M_\odot$ so as to have evolved off the main sequence in less time than the age of the stellar population (<5 Gyr). This tidal stripping mechanism has been invoked to explain the hot stars in the Galactic Centre[24], and the rate of tidal disruption of tidally stripped stars is likely to be higher than that of solar-type stars[25]. The mass of the black hole derived from the light-curve fit depends on the mass and radius of the star at the time of disruption. Using $M_*\approx0.23M_\odot$ and $R_*\approx0.33R_\odot$ (values measured for a red giant core that was stripped in a binary system[26]), and assuming that the evolution of the core is similar to one that is tidally stripped, we find that $f=M_{\rm acc}/M_*\gtrsim0.058$ (approaching $f\gtrsim0.1$ as measured in simulations[27]), that $M_{\rm BH}=(2.8\pm0.1)\times10^6M_\odot$ and that the peak luminosity approaches the Eddington luminosity of the supermassive black hole ($L_{\rm peak}\gtrsim0.6L_{\rm Edd}$).

Received 8 November 2011; accepted 23 February 2012.
Published online 2 May 2012.

1. Rees, M. J. Tidal disruption of stars by black holes of 10 to the 6th-10 to the 8th solar masses in nearby galaxies. *Nature* **333,** 523–528 (1988).
2. Komossa, S. & Bade, N. The giant X-ray outbursts in NGC 5905 and IC 3599: follow-up observations and outburst scenarios. *Astron. Astrophys.* **343,** 775–787 (1999).
3. Komossa, S. *et al.* A huge drop in the X-ray luminosity of the nonactive galaxy RX J1242.6–1119A, and the first postflare spectrum: testing the tidal disruption scenario. *Astrophys. J.* **603,** L17–L20 (2004).
4. Esquej, P. *et al.* Evolution of tidal disruption candidates discovered by XMM-Newton. *Astron. Astrophys.* **489,** 543–554 (2008).
5. Gezari, S. *et al.* Luminous thermal flares from quiescent supermassive black holes. *Astrophys. J.* **698,** 1367–1379 (2009).
6. van Velzen, S. *et al.* Optical discovery of probable stellar tidal disruption flares. *Astrophys. J.* **741,** 73–96 (2011).
7. Bloom, J. S. *et al.* A possible relativistic jetted outburst from a massive black hole fed by a tidally disrupted star. *Science* **333,** 203–206 (2011).
8. Burrows, D. N. *et al.* Relativistic jet activity from the tidal disruption of a star by a massive black hole. *Nature* **476,** 421–424 (2011).
9. Zauderer, B. A. *et al.* Birth of a relativistic outflow in the unusual γ-ray transient Swift J164449.3+573451. *Nature* **476,** 425–428 (2011).
10. Cenko, S. B. *et al.* Swift J2058.4+0516: discovery of a possible second relativistic tidal disruption flare. Preprint at ⟨http://arxiv.org/abs/1107.5307⟩ (2011).
11. Phinney, E. S. in *The Center of the Galaxy* (ed. Morris, M.) 543–553 (IAU Symp. 136, Kluwer, 1989).
12. Evans, C. R. & Kochanek, C. S. The tidal disruption of a star by a massive black hole. *Astrophys. J.* **346,** L13–L16 (1989).
13. Ulmer, A. Flares from the tidal disruption of stars by massive black holes. *Astrophys. J.* **514,** 180–187 (1999).
14. Kaiser, N. *et al.* The Pan-STARRS wide-field optical/NIR imaging survey. *Proc. SPIE* **7733,** 77330E (2010).
15. Martin, D. C. *et al.* The Galaxy Evolution Explorer: a space ultraviolet survey mission. *Astrophys. J.* **619,** L1–L6 (2005).
16. Aihara, H. *et al.* The eighth data release of the Sloan Digital Sky Survey: first data from SDSS-III. *Astrophys. J. Suppl. Ser.* **193,** 29–45 (2011).
17. Lawrence, A. *et al.* The UKIRT Infrared Deep Sky Survey (UKIDSS). *Mon. Not. R. Astron. Soc.* **379,** 1599–1617 (2007).
18. Blanton, M. R. & Roweis, S. K-corrections and filter transformations in the ultraviolet, optical, and near-infrared. *Astron. J.* **133,** 734–754 (2007).
19. Häring, N. & Rix, H.-W. On the black hole mass-bulge mass relation. *Astrophys. J.* **604,** L89–L92 (2004).
20. Lodato, G., King, A. R. & Pringle, J. E. Stellar disruption by a supermassive black hole: is the light curve really proportional to $t^{-5/3}$? *Mon. Not. R. Astron. Soc.* **392,** 332–340 (2009).
21. Strubbe, L. E. & Quataert, E. Optical flares from the tidal disruption of stars by massive black holes. *Mon. Not. R. Astron. Soc.* **400,** 2070–2084 (2009).
22. Lawrence, A. The UV peak in active galactic nuclei: a false continuum from blurred reflection? Preprint ⟨http://arxiv.org/abs/1110.0854⟩ (2011).
23. Loeb, A. & Ulmer, A. Optical appearance of the debris of a star disrupted by a massive black hole. *Astrophys. J.* **489,** 573–578 (1997).
24. Davies, M. B. & King, A. The stars of the galactic center. *Astrophys. J.* **624,** L25–L27 (2005).
25. Kobayashi, S., Laguna, P., Phinney, E. S. & Mészáros, P. Gravitational waves and X-ray signals from stellar disruption by a massive black hole. *Astrophys. J.* **615,** 855–865 (2004).
26. Maxted, P. F. L. *et al.* Discovery of a stripped red giant core in a bright eclipsing binary system. *Mon. Not. R. Astron. Soc.* **418,** 1156–1164 (2011).
27. Ayal, S., Livio, M. & Piran, T. Tidal disruption of a solar-type star by a supermassive black hole. *Astrophys. J.* **545,** 772–780 (2000).
28. Steffen, A. T. *et al.* The X-ray-to-optical properties of optically selected active galaxies over wide luminosity and redshift ranges. *Astron. J.* **131,** 2826–2842 (2006).
29. Steele, I. A. *et al.* The Liverpool Telescope: performance and first results. *Proc. SPIE* **5389,** 679 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Tananbaum for approving our Chandra Director's Discretionary Time request. We are grateful to G. Lodato for providing the tidal disruption event models in tabular form, and to S. Moran for running software to calculate the host-galaxy K-corrections. We thank R. E. Williams for discussions on the line emission in the spectra. S.G. was supported by NASA through a Hubble Fellowship grant awarded by the Space Telescope Science Institute, which is operated by AURA Inc. for NASA. Partial support for this work was provided by the National Science Foundation. The PS1 survey has been made possible through contributions of the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, The Johns Hopkins University, Durham University, the University of Edinburgh, Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Inc. and the National Central University of Taiwan, and by NASA under a grant issued through the Planetary Science Division of the NASA Science Mission Directorate. We acknowledge NASA's support for construction, operation, and science analysis of the GALEX mission, which was developed in cooperation with Centre National d'Etudes Spatiales of France and the Korean Ministry of Science and Technology. Some of the observations reported here were obtained at the MMT Observatory, which is a joint facility of the Smithsonian Institution and the University of Arizona, and at the Liverpool Telescope, which is operated with financial support from the UK Science and Technology Facilities Council. The computations in this paper were run on the Odyssey cluster supported by the FAS Science Division Research Computing Group at Harvard University. R.J.F. is a Clay Fellow.

Author Contributions S.G. designed the observations and the transient detection pipeline for the GALEX TDS, and measured the ultraviolet photometry of PS1-10jh. K.F.

and J.D.N coordinated, and D.C.M. facilitated, the GALEX TDS observations. A.R. designed the PhotPipe transient detection pipeline hosted by Harvard/CfA for the PS1 Medium Deep Survey (MDS), and measured the optical photometry of PS1-10jh. R.C. designed, implemented and analysed the MMT optical spectroscopy observations, and contributed to the operation of PhotPipe and the visual inspection of transient alerts. E.B. proposed and facilitated the MMT observations. M.E.H., G.N., D.S. and R.J.F. contributed to the operation of PhotPipe and the visual inspection of transient alerts. P.J.C., R.J.F., G.H.M., L.C. and A.S. contributed to the MMT observations. S.J.S. designed, and K.S. operated, the transient pipeline for PS1 MDS hosted by Queen's University Belfast. C.W.S., J.L.T. and W.M.W.-V. facilitated the transient pipelines for PS1 MDS. W.S.B., K.C.C., T.G., J.N.H., N.K., R.-P.K., E.A.M., J.S.M., P.A.P., C.W.S. and J.L.T. helped build the PS1 system. S.G. requested the Director's Discretionary Time Chandra X-ray observation and analysed the data. A.L. obtained the Liverpool Telescope optical imaging observations and analysed the data, and stimulated discussions on the nature of the SED of PS1-10jh. S.G. analysed and modelled the multicolour light curve and the SED of PS1-10jh. T.H. and C.N. stimulated discussions on the nature of the disrupted star. The paper was organized and written by S.G., and all authors provided feedback on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.G. (suvi@pha.jhu.edu).

LETTER

doi:10.1038/nature11002

Patterning by controlled cracking

Koo Hyun Nam[1], Il H. Park[1] & Seung Hwan Ko[2]

Crack formation drives material failure and is often regarded as a process to be avoided[1–3]. However, closer examination of cracking phenomena has revealed exquisitely intricate patterns such as spirals[4], oscillating[5,6,7] and branched[7] fracture paths and fractal geometries[8]. Here we demonstrate the controlled initiation, propagation and termination of a variety of channelled crack patterns in a film/substrate system[9–11] comprising a silicon nitride thin film deposited on a silicon substrate using low-pressure chemical vapour deposition. Micro-notches etched into the silicon substrate concentrated stress for crack initiation, which occurred spontaneously during deposition of the silicon nitride layer. We reproducibly created three distinct crack morphologies—straight, oscillatory and orderly bifurcated (stitchlike)—through careful selection of processing conditions and parameters. We induced direction changes by changing the system parameters, and we terminated propagation at preformed multi-step crack stops. We believe that our patterning technique presents new opportunities in nanofabrication and offers a starting point for atomic-scale pattern formation[12], which would be difficult even with current state-of-the-art nanofabrication methodologies.

We have found that the tools and techniques for thin brittle material deposition drawn from conventional microfabrication technologies allow us to attain nearly ideal conditions for crack formation[13,14]. Low-pressure chemical vapour deposition of silicon nitride (Si_3N_4) thin films offers high levels of uniformity, reproducibility on planar surfaces and superior controllability for batch processing. A well refined, single-crystalline silicon substrate gives predictable thermal responses. Finally, standard silicon wafer and microfabrication techniques enable the control of processing conditions and facilitate post-fabrication processes. To control crack initiation at the desired position and orientation at microscopic scales, we used silicon etching to create 'micro-notch' structures which concentrate stress to initiate cracks. With uniform deposition of a cracking layer on the substrate, this technique can reduce undesirable effects, and permit the generation of cracks that are much more complex and controllable than those generated on as-fabricated samples. In the present case, the crack occurs on the thin Si_3N_4 layer during deposition, as a result of the film stress arising between that layer and the underlying material. Figure 1a shows the controlled formation of oscillating cracks on a Si_3N_4 thin film deposited on a silicon substrate with patterned notches. The geometric characteristics of the cracks, such as their wavelength and amplitude of oscillation, are influenced by the adjacent stress field, which is defined by the value of the film stress at the time of cracking. Thus, successful crack initiation can be achieved with a micro-notch designed with an optimal notch tip angle to concentrate stress (Supplementary Fig. 3). The characteristics of the notch determine the timing of crack initiation and, in turn, the condition of the substrate and cracking material where the propagation occurs. Thus, the design of a micro-notch changes the characteristics of the resulting cracks. We were able to achieve highly controllable crack initiation with 100% yield (Fig. 1a) through successful micro-notch design.

Because the orientations and shapes of cracks in the thin layer atop the silicon substrate are determined by the film stress occurring between two different materials, variations in the types and thicknesses of testing materials were thought to be significant factors that could alter the cracking response. Oscillating and straight cracks (Fig. 1b, c; left and middle images) are generated on a Si_3N_4 thin film deposited on silicon wafers of different crystallographic orientations (Supplementary Fig. 4) under controlled conditions. Both of those types of crack can coexist on a (100) silicon wafer, and in that case, they meet each other perpendicularly, as shown in Fig. 1d. Crack widths observed in this study varied between about 10 and 120 nm, with straight cracks generally narrower than oscillating cracks. Straight cracks are also observed when a silicon dioxide (SiO_2) film is deposited as an interlayer between the Si_3N_4 and the silicon substrate. In this circumstance, 'stitchlike' cracks, with non-propagating branches[3], can also form (Fig. 1b, c; right images). The non-propagating branches are caused by fluctuations in the available elastic energy, above and below the value required for branching. Unlike the straight cracks, the stitchlike cracks form adjacent to a pre-existing underlying crack in the SiO_2 interlayer. The highly ordered branching observed in the stitchlike cracks is unusual, and testifies to the very precisely controlled conditions of our study; crack bifurcation occurs at high energies, where instability is also more readily introduced[3,15]. The precise control of crack formation by micro-notching can yield predefined complex nano-patterning, as shown in Fig. 1e and Supplementary Fig. 5.

The three main types of crack can be created by controlling the processing conditions and system parameters, and by using the intrinsic crystallographic properties of the silicon substrate. Oscillating cracks are quite constant in profile along their length, and have a strong tendency to propagate in the <110> direction through a (100) silicon wafer (Fig. 2a). Our oscillating cracks resemble those reported in refs 7 and 16, but are easier to create and can be reliably replicated. They should also be distinguished from previously studied 'through-the-thickness' cracks[3–8,14–16], which have frequently been used to study crack dynamics. In the present case, we have a two-dimensional, channelling crack that propagates towards, and through, the interface between two different materials, which is additionally characterized by an in-plane propagation across each of the materials[9–11] (Fig. 2b; and see Supplementary Information for further discussion of the physics of the oriented cracking).

When a crack's depth of penetration in the film/substrate system is shallow, it may propagate in several specific directions, owing to the different stress distributions in the film and the substrate[17]. For example, a straight crack shows several propagation directions in the vicinity of the <100> orientation. However, when the crack's depth of penetration into the silicon wafer is increased, the crack propagation direction is more strongly influenced by the crystalline orientation of the substrate than by the anisotropically oriented stress distribution in the film/substrate system. We have found that if a propagating crack experiences a substantial change in either of these conditions, its propagation direction changes, in a manner resembling the refraction of light. As shown in Fig. 3, a change in the underlying material initially causes kinking[18,19] of the crack when it passes across the interface between the two different regions, and then maintenance of the new propagation direction. In the 'interlayer region', formed of a three-layer

[1]Research Center of MEMS Space Telescope, Department of Physics, Ewha Womans University, Daehyun-dong 11-1, Seodaemun-gu, Seoul 120-750, South Korea. [2]Applied Nano Technology and Science Lab, Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, 335 Daehak-ro, Yusong-Gu, Daejeon 305-701, South Korea.

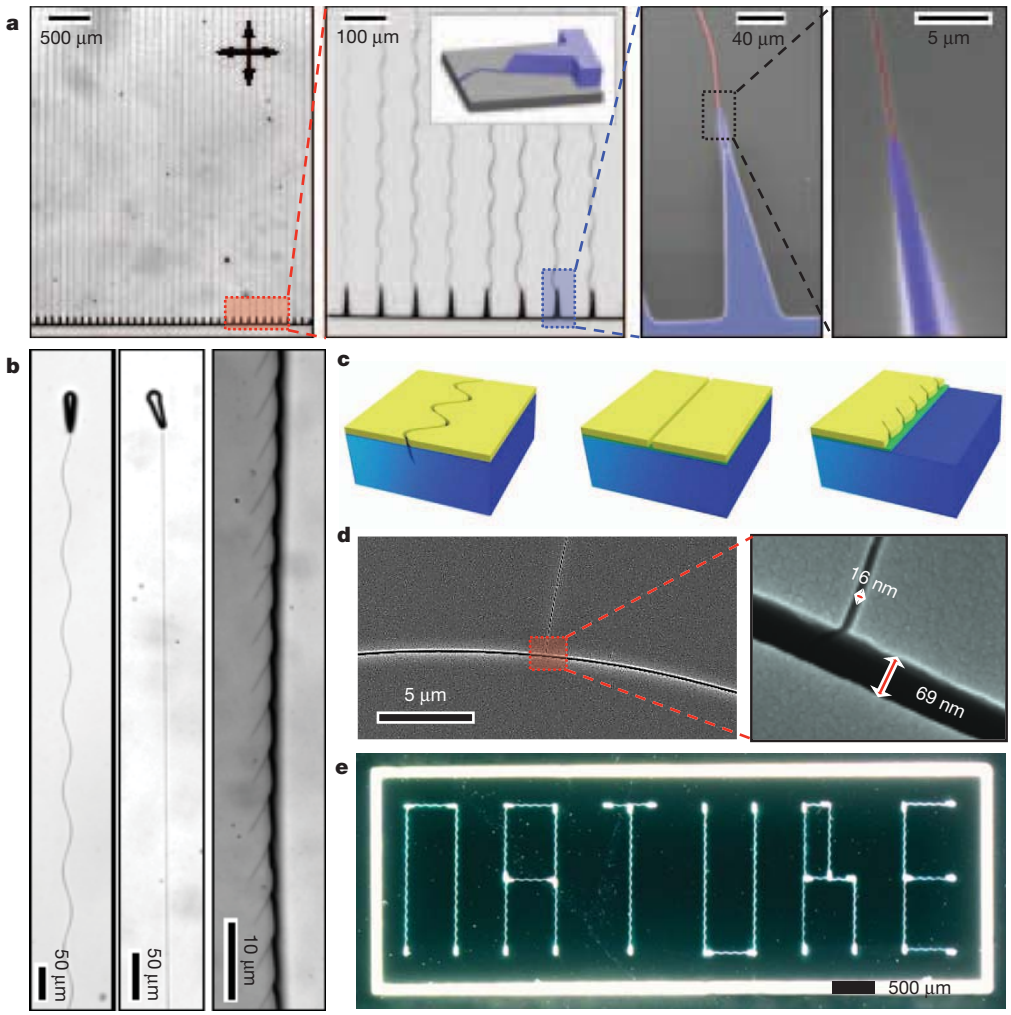


Figure 1 | Crack control. **a**, Optical and electron micrographs of oscillating crack initiation from micro-notches in a Si_3N_4 thin film deposited on (100) silicon. **b**, Left to right: an oscillating crack initiated from a micro-notch; a straight crack initiated from a micro-notch; and a stitchlike crack showing orderly bifurcated cracking with non-propagating branches. **c**, Schematic diagrams of each type of crack. Yellow, Si_3N_4 thin film; cyan, SiO_2 interlayer; blue, silicon wafer. **d**, Electron micrographs of cracks of different types and widths. The straight crack propagating vertically in the figure has much a smaller width than the oscillating crack running horizontally. **e**, Precise manipulation of oscillating cracks with crack notches to write the word 'NATURE'.

composite, or trimaterial[20,21], a SiO_2 interlayer acts as a buffer layer, suppressing the penetration of the crack into the silicon substrate. In the 'no-interlayer region', where the SiO_2 interlayer has been selectively etched away, the crack penetrates deeply into the substrate. (See Supplementary Information for further discussion of the crack refraction mechanism.) The cracks require an adequate distance of travel through these transition regions to establish the new mode of crack propagation. If the crack propagation distance is short, crack kinking is observed but not a stable straight or oscillatory mode (Fig. 3b).

Arresting crack propagation is important for the prevention of material failure, and the tailoring of controlled cracks can find use in engineered structures. As has been detailed in previous studies[22–25], several methods for limiting cracks have been developed, both through the addition of reinforcing materials and through the inclusion of

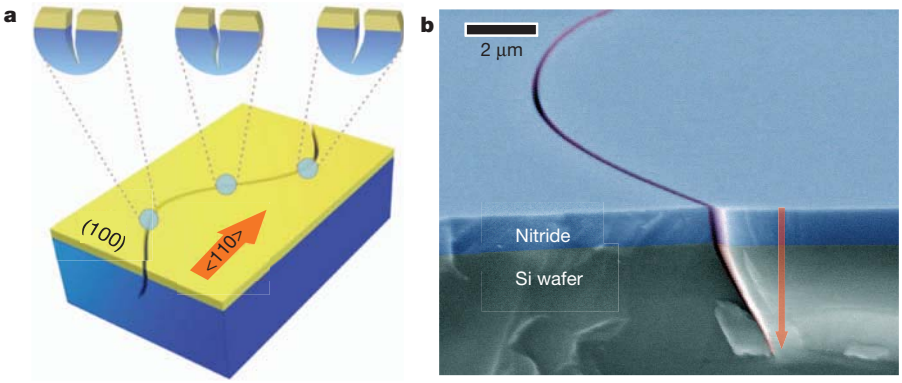


Figure 2 | Formation mechanism of an oscillating crack. **a**, Schematic figure of cracking materials and crack propagation, showing crystallographic orientations of the silicon substrate. Oscillating cracks tend to propagate in the <110> direction, and also penetrate downwards into the silicon substrate. These in-substrate cracks point towards the centre of the crack trajectory, displaying maximal angles from the vertical along the substrate's lowest energy cleavage plane, {111}, at the points farthest from the oscillating centre (as shown in the cross-sections above the plane). The penetration depth of the crack into the substrate depends, in part, on the substrate's bimaterial bending force, and, in turn, determines the amplitude of oscillation. (See Supplementary Information for further discussion.) **b**, Electron micrograph showing penetration of a crack into the silicon substrate (red arrow), angling towards the propagating centre of oscillation.

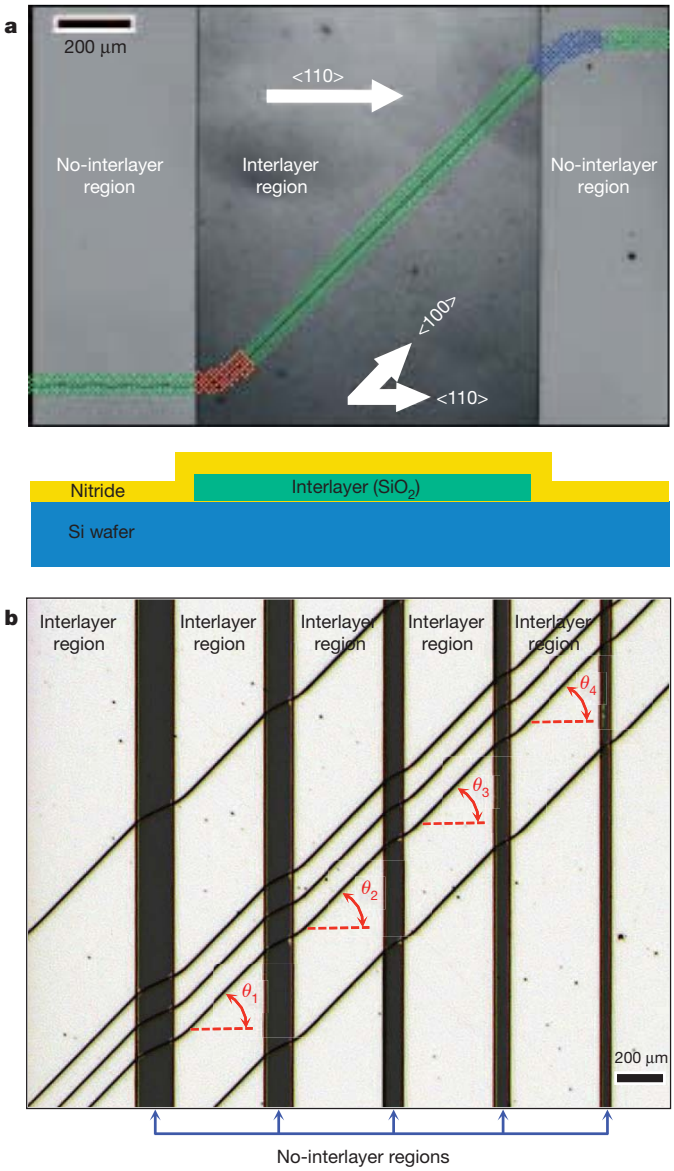


Figure 3 | Crack refraction. a, Cracks running through a Si_3N_4 film on a SiO_2 interlayer tend to propagate straight in the <100> direction, whereas oscillating cracks run in the <110> direction. (See also Supplementary Fig. 4.) When a crack passes across the interface between two different underlying configurations (shown in the schematic cross-section below), the crack trajectory kinks[18,19], until it attains the propagation direction appropriate to the configuration of the region through which it is then travelling. Two such 'transition regions' are highlighted in red and blue. **b**, Crack propagation across multiple alternating regions in which the no-interlayer regions are insufficiently wide to allow a full oscillation cycle to occur. Straight cracks in the interlayer regions, however, have sufficient distance to recover identical propagation angles ($\theta_1 = \theta_2 = \theta_3 = \theta_4$).

impurities to increase fracture resistance when a physical separation of the cracking materials is not possible. However, these methods introduce a material preparation step into the crack prevention process, and they are unable to precisely place crack stops at specific locations. In a film/substrate composite, varying the local stress in the film by altering the substrate geometry could, to an extent, affect the dynamics of cracks. However, in our cases the film continuity is barely affected by the structural geometry of the underlying materials, including the silicon substrate, because the Si_3N_4 thin film prepared by chemical vapour deposition is highly conformal. Because the stress field between the substrate and deposited film in this case is uniform, a crack will not stop propagating until it has reached the edge of the wafer.

If the stress drops during the propagation of a crack, the crack registers the disturbance of the driving force. When the driving force decreases to a value comparable to the crack resistance, the crack stops propagating. We attempted to decrease the driving force by fabricating a stair-profiled structure on the substrate. Fig. 4a, b shows a substrate patterned with terraced edges by deep reactive-ion etching; in these regions, the film stress drops sharply in front of the propagating crack. However, when overstress is present in the film, the surplus energy that remains after the consumption of the elastic energy is stored by the enhanced acceleration of a crack or extension of the substrate penetration. This excess energy is the major obstacle in the effort to terminate crack propagation. We have found that a significant change in the stress field to stop crack propagation is difficult in normal one-step structures fabricated by conventional etching processes (Fig. 4a). In this case, the crack recovers its driving force from the stored energy, which is then regenerated during the subsequent propagation. On the other hand (Fig. 4b), crack propagation terminates when multi-step structures are introduced to reduce the propagation distance to a level inadequate for the recovery of surplus energy within the travel region, at which point the stored energy becomes insufficient for further propagation of the crack. As shown in the left inset of Fig. 4b, we fabricated a stair profile with multiple steps of 5 μm height and width, and found that this structure is able to stop crack propagation with no failures (Fig. 4b, d), whereas a crack stop with a single step profile fails to terminate crack propagation (Fig. 4a, c). As such a stair structure is difficult to construct using conventional microfabrication processes, we invented a special single-step lithography method to fabricate stair-profile microstructures by intentional diffraction (see Supplementary Fig. 2 for process details). In addition to providing stress control, stair-profile microstructures are also able to disrupt the stress field by increasing the roughness of the etched surface, which in turn obstructs crack propagation. Another important feature of the crack stop is its ability to protect sampling regions from the intrusion of highly uncontrollable external cracks, which develop readily during the wafer dicing process (Fig. 4d). (See Supplementary Information for further discussion of the crack stoppage mechanism.) An isolated region enclosed and protected by a crack stop remains crack-free even after wafer dicing.

The precise control of crack initiation and termination by the proposed crack notches and stops may enable the development of a new, very simple approach for high-resolution, arbitrary nano-patterning of large-area substrates, as a potential alternative to such state-of-the-art approaches as electron-beam or high-energy beam lithography[26,27]. Large-area, high-resolution nano-patterning is challenging even with conventional electron-beam lithography, which is a very expensive, time-consuming and slow serial process[26,27]. By contrast, our approach is not limited by wafer size and can be easily scaled up to much larger wafer sizes with no increase in processing time and cost, owing to the parallel and self-generated process used in production. Furthermore, since the materials and fabrication processes used in this study are fully compatible with well developed, silicon-based integrated circuit processes, this technique should be readily applicable in the semiconductor industry, where the level of scale-down determines competitiveness. (See Supplementary Information for further discussion.)

So far, we have focused on an optimum combination of materials (Si_3N_4 thin films on silicon wafers) for the study of controlled cracking. Other material combinations might satisfy the criteria for triggering crystallographically oriented channelling cracks (see Supplementary Information). We expect that more precise control of dimensions, including crack width and depth, will be achieved through post-processing, such as additional conformal deposition[28] and polishing. Finally, more elaborate control of the experimental environment than has yet been achieved will open up new scientific insights and technological consequences of this phenomenon[1,2,13,14,29,30]. (See Supplementary Fig. 11 for more advanced crack manipulation possibilities.)

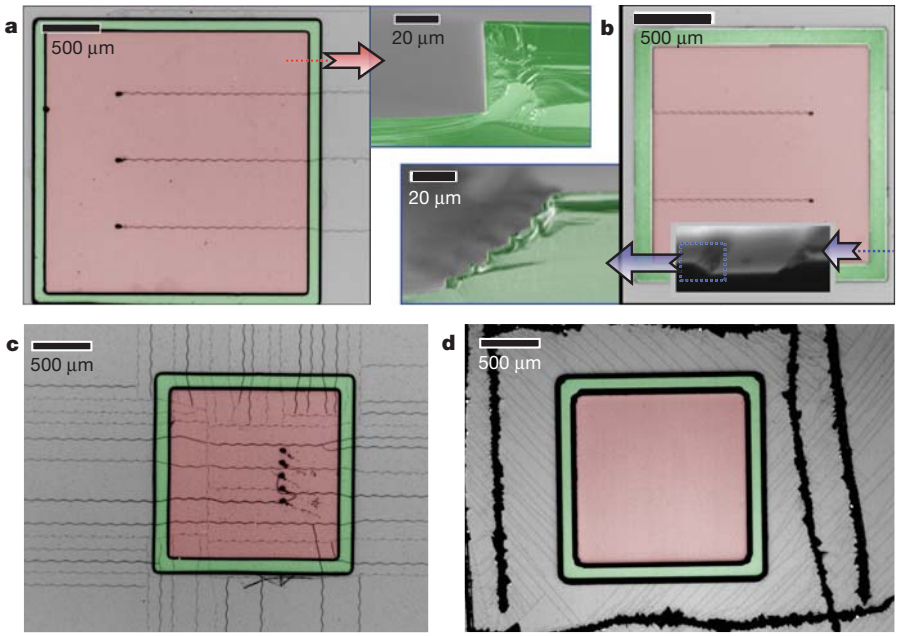


Figure 4 | Crack termination. **a**, Unsuccessful crack arrest. A box structure (green colour) intended for crack arrest is built using a one-step structure generated by conventional lithography. On the conformably deposited cracking layer, all cracks pass through the one-step structure (shown in cross-section in the electron micrograph at right) without stopping. **b**, Successful crack arrest. The same box structure, but with a multi-step edge fabricated by our new technique of diffraction-induced stairing lithography (DISL; Supplementary Fig. 2). In this case, all cracks terminate when they encounter the box structure (shown in cross-section in the electron micrograph at left). **c**, **d**, Additional examples of crack stoppage failure (**c**) and success (**d**). Structures etched in the silicon substrate without DISL influence the propagation of cracks to some extent, but are unable to cause termination.

METHODS SUMMARY

Samples used in this study were made mostly from 525-μm-thick (100) silicon wafers, with (110) and (111) silicon samples used to observe the dependence of crack propagation on crystallographic orientation in silicon substrates. Si_3N_4 (stoichiometric silicon nitride) films were formed by thin-film deposition from a chemical precursor gas in a low-pressure environment at accurately controlled temperature (800 °C) and pressure (200 mtorr). Source gases and their mass flow rates were as follows: dichlorosilane (H_2SiCl_2) at 30 cm^3 STP min^{-1} and ammonia (NH_3) at 100 cm^3 STP min^{-1}. SiO_2 interlayers were deposited at 1,000 °C by thermal oxidation, and we attempted various deposition thicknesses between 100 nm and 2 μm. The fabrication method for the crack initiation notch was designed to be an easy process involving few steps, and consequently variations in the structure of the notches were minimized. Although the height of a notch is determined by other structures fabricated together with the notch, the possibility of crack initiation at the notch reaches a satisfactory level when the height is more than 5 μm. For crack refraction, a SiO_2 interlayer was defined by chemical etching, and Si_3N_4 was deposited on the interlayer using the chemical thin film deposition procedure mentioned earlier. For details of the microfabrication process, see Supplementary Information.

Received 21 November 2011; accepted 28 February 2012.

1. Hellemans, A. Cracks: more than just a clean break. *Science* **281,** 943–944 (1998).
2. Livne, A., Bouchbinder, E., Svetlizky, I. & Fineberg, J. The near-tip fields of fast cracks. *Science* **327,** 1359–1363 (2010).
3. Broek, D. *Elementary Engineering Fracture Mechanics* 4th edn (Martinus Nijhoff Publishers, 1986).
4. Leung, K.-T., Jozsa, L., Ravasz, M. & Neda, Z. Pattern formation: spiral cracks without twisting. *Nature* **410,** 166 (2001).
5. Deegan, R. D., Petersan, P. J., Marder, M. & Swinney, H. L. Oscillating fracture paths in rubber. *Phys. Rev. Lett.* **88,** 014304 (2001).
6. Deegan, R. D. *et al.* Wavy and rough cracks in silicon. *Phys. Rev. E* **67,** 066209 (2003).
7. Yuse, A. & Sano, M. Transition between crack patterns in quenched glass plates. *Nature* **362,** 329–331 (1993).
8. Skjeltorp, A. T. & Meakin, P. Fracture in microsphere monolayers studied by experiment and computer simulation. *Nature* **335,** 424–426 (1988).
9. Hutchinson, J. W. & Suo, Z. Mixed mode cracking in layered materials. *Adv. Appl. Mech.* **29,** 63–191 (1991).
10. Beuth, J. L. Jr. Cracking of thin bonded films in residual tension. *Int. J. Solids Struct.* **29,** 1657–1675 (1992).
11. Ye, T., Suo, Z. & Evans, A. G. Thin film cracking and the roles of substrate and interface. *Int. J. Solids Struct.* **29,** 2639–2648 (1992).
12. Ball, P. *Nature's Patterns: a Tapestry in Three Parts* (Oxford Univ. Press, 2009).
13. Marder, M. & Fineberg, J. How things break. *Phys. Today* **49,** 24–29 (1996).
14. Bouchbinder, E., Fineberg, J. & Marder, M. Dynamics of simple cracks. *Annu. Rev. Cond. Matter Phys.* **1,** 371–395 (2010).
15. Sharon, E., Gross, S. P. & Fineberg, J. Local crack branching as a mechanism for instability in dynamic fracture. *Phys. Rev. Lett.* **74,** 5096–5099 (1995).
16. Yuse, A. & Sano, M. Instabilities of quasi-static crack patterns in quenched glass plates. *Physica D* **108,** 365–378 (1997).
17. Kobeda, E. & Irene, E. A. Intrinsic SiO_2 film stress measurements on thermally oxidized Si. *J. Vac. Sci. Technol. B* **5,** 15–19 (1987).
18. Cotterell, B. & Rice, J. R. Slightly curved or kinked cracks. *Int. J. Fract.* **16,** 155–169 (1980).
19. He, M.-Y. & Hutchison, J. W. Kinking of a crack out of an interface. *J. Appl. Mech.* **56,** 270–278 (1989).
20. Choi, S. T. & Earmme, Y. Y. Elastic study on singularities interacting with interfaces using alternating technique: Part I. Anisotropic trimaterial. *Int. J. Solids Struct.* **39,** 943–957 (2002).
21. Choi, S. T. & Earmme, Y. Y. Elastic study on singularities interacting with interfaces using alternating technique: Part II. Isotropic trimaterial. *Int. J. Solids Struct.* **39,** 1199–1211 (2002).
22. Faber, K. T. & Evans, A. G. Crack deflection processes – I. Theory. *Acta Metall.* **31,** 565–576 (1983).
23. Green, D. J., Tandon, R. & Sglavo, V. M. Crack arrest and multiple cracking in glass through the use of designed residual stress profiles. *Science* **283,** 1295–1297 (1999).
24. Rao, M. P., Sánchez-Herencia, A. J., Beltz, G. E., McMeeking, R. M. & Lange, F. F. Laminar ceramics that exhibit a threshold strength. *Science* **286,** 102–105 (1999).
25. Clegg, W. J. Controlling cracks in ceramics. *Science* **286,** 1097–1099 (1999).
26. Son, Y. *et al.* Nanoscale electronics: digital fabrication by direct femtosecond laser processing of metal nanoparticles. *Adv. Mater.* **23,** 3176–3181 (2011).
27. Xia, Y. *et al.* One dimensional nanostructures: synthesis, characterization, and applications. *Adv. Mater.* **15,** 353–389 (2003).
28. Nam, S. W. *et al.* Sub-10-nm nanochannels by self-sealing and self-limiting atomic layer deposition. *Nano Lett.* **10,** 3324–3329 (2010).
29. Marder, M. Cracks takes a new turn. *Nature* **362,** 295–296 (1993).
30. Buehler, M. J. & Gao, H. Dynamical fracture instabilities due to local hyperelasticity at crack tips. *Nature* **439,** 307–310 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by Creative Research Initiatives (Research Center of MEMS Space Telescope) of MEST/NRF. We thank Y.Y. Earmme at KAIST for discussions and J. Yeo, Y. D. Suh, S. Hong, P. Lee, Y. Rho and J.-A. Jeon for technical assistance with fabrications and experiments.

Author Contributions K.H.N. conceived the study, discovered the control of cracking using microfabrication, conducted experiments and theoretical study of the phenomena, and invented DISL. K.H.N. and I.H.P. designed mask patterns for photolithography and fabricated samples. S.H.K. did the post-processing and conducted experiments. K.H.N. and S.H.K. wrote the paper and discussed the results. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.H.N. (koonam@namk.org) or S.H.K. (maxko@kaist.ac.kr).

LETTER

doi:10.1038/nature11064

Twenty-first-century warming of a large Antarctic ice-shelf cavity by a redirected coastal current

Hartmut H. Hellmer[1], Frank Kauker[1,2], Ralph Timmermann[1], Jürgen Determann[1] & Jamie Rae[3]

The Antarctic ice sheet loses mass at its fringes bordering the Southern Ocean. At this boundary, warm circumpolar water can override the continental slope front, reaching the grounding line[1,2] through submarine glacial troughs and causing high rates of melting at the deep ice-shelf bases[3,4]. The interplay between ocean currents and continental bathymetry is therefore likely to influence future rates of ice-mass loss. Here we show that a redirection of the coastal current into the Filchner Trough and underneath the Filchner–Ronne Ice Shelf during the second half of the twenty-first century would lead to increased movement of warm waters into the deep southern ice-shelf cavity. Water temperatures in the cavity would increase by more than 2 degrees Celsius and boost average basal melting from 0.2 metres, or 82 billion tonnes, per year to almost 4 metres, or 1,600 billion tonnes, per year. Our results, which are based on the output of a coupled ice–ocean model forced by a range of atmospheric outputs from the HadCM3[5] climate model, suggest that the changes would be caused primarily by an increase in ocean surface stress in the southeastern Weddell Sea due to thinning of the formerly consolidated sea-ice cover. The projected ice loss at the base of the Filchner–Ronne Ice Shelf represents 80 per cent of the present Antarctic surface mass balance[6]. Thus, the quantification of basal mass loss under changing climate conditions is important for projections regarding the dynamics of Antarctic ice streams and ice shelves, and global sea level rise.

The Weddell Sea (Fig. 1) is dominated by a cyclonic gyre circulation that allows Circumpolar Deep Water to enter only from the east[7]. Within the southern branch of the gyre, the water mass can be identified as the Weddell Sea's temperature maximum at a depth of ~300 m. The temperature decreases from 0.9 °C at the Greenwich meridian[7] to 0.6 °C off the tip of the Antarctic Peninsula[8]. Only traces of the relatively warm water penetrate the broad southern continental shelf[9], reaching the Filchner–Ronne Ice Shelf front with a temperature of −1.5 °C (ref. 10). However, no indications exist that this water mass advances far into the ice-shelf cavity[11]. Instead, locally formed high-salinity shelf water at the surface freezing temperature (about −1.89 °C) fuels a sub-ice-shelf circulation that brings the heat to the deep southern grounding line, where the base of the ice shelf touches the ground. High-salinity shelf water is the densest water mass in the Weddell Sea, and is formed by brine rejection during sea-ice formation on a southward-sloping continental shelf. The need for a dense water mass to transport heat to the grounding line was used as an argument for the Filchner–Ronne Ice Shelf to be protected in a warmer climate[12]. This hypothesis assumes that rising atmospheric temperatures reduce sea-ice formation and, thus, the densification of the shelf water masses. This view considers solely the formation of dense continental shelf water masses in a warmer climate, but less-consolidated sea-ice cover might also influence the Weddell Sea circulation, including the course of the coastal current.

The marine-based West Antarctic Ice Sheet has the potential to contribute 3.3 m to the global eustatic sea-level rise[13]. Its ice shelves fringing the Amundsen Sea are exposed today to Circumpolar Deep Water with temperatures of more than 1 °C. This water mass cascades nearly undiluted from the continental shelf break into ~1,000-m-deep trenches underlying the floating extensions of ice streams that drain the West Antarctic Ice Sheet[14]. Some ice streams from this ice sheet also feed the 449,000-km^2 Filchner–Ronne Ice Shelf (Fig. 1), forming the southern coast of the Weddell Sea. These ice streams pass over mountain ranges and thus would not face an increase in basal melting as the grounding line retreats. However, major ice streams entering the Filchner–Ronne Ice Shelf discharge large catchment basins of the East Antarctic Ice Sheet[15]. Once afloat, this ice interacts with the waters of the Weddell Sea.

We forced the Bremerhaven Regional Ice–Ocean Simulations (BRIOS) model[16] with the atmospheric output of two versions of the HadCM3 climate model (Table 1). Whereas HadCM3-A is the baseline simulation used in perturbed physics ensembles[17], HadCM3-B is a model configuration with an interactive carbon cycle and vegetation, and is used in the ENSEMBLES project[18]. We used the output of two simulations of the twentieth century (HadCM3-A (1900–1999) and HadCM3-B (1860–1999)) and the Intergovernmental Panel on Climate Change scenarios E1 (2000–2199)[19] and A1B (2000–2099/2199)[20] (Table 1). These scenarios are characterized by different carbon dioxide emissions, with atmospheric concentrations reaching

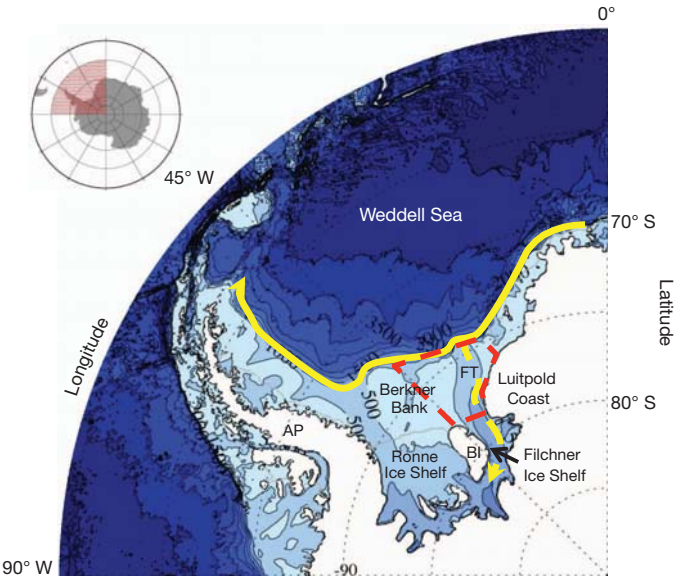


Figure 1 | Map of Weddell Sea bathymetry south of 60° S. Bathymetry is based on RTopo-1 (ref. 29) with a colour contour interval of 500 m. Inset represents the model domain, with the red dashed line showing the map location within the circumpolar Southern Ocean. The solid yellow arrow marks the present course of the coastal current in the Weddell Sea. The possibility of pulsing into the Filchner Trough (FT) is marked by the dashed yellow arrow. The region bounded by the dashed red line provided the integrated and mean values in Fig. 3. The solid grey line off the coastline indicates the ice-shelf front. AP, Antarctic Peninsula; BI, Berkner Island.

[1]Alfred Wegener Institute for Polar and Marine Research, D-27570 Bremerhaven, Germany. [2]OASys, Lerchenstrasse 28a, 22767 Hamburg, Germany. [3]Met Office Hadley Centre, Exeter EX1 3PB, UK.

Table 1 | List of BRIOS model experiments

Model	Simulation	Period
HadCM3-A	20th century	1900–1999
HadCM3-A	A1B	2000–2099
HadCM3-B	20th century	1860–1999
HadCM3-B	A1B	2000–2199
HadCM3-B	E1	2000–2199

Atmospheric forcing was extracted from the results of the climate models HadCM3-A and HadCM3-B. HadCM3-A forcing extends only until 2099 and is not available for the scenario E1. Scenarios E1 and A1B are characterized by different carbon dioxide emissions, with atmospheric concentrations reaching 450 parts per million by volume (p.p.m.v.) and 700 p.p.m.v. by the year 2100, respectively.

450 p.p.m.v. and 700 p.p.m.v. by the year 2100, respectively. BRIOS is a coupled ice–ocean model that resolves the Southern Ocean at latitudes south of $\phi = 50^\circ$ S zonally with a resolution of 1.5° and meridionally with a resolution of $1.5^\circ \times \cos(\phi)$. The water column is variably divided into 24 terrain-following layers. The sea-ice component is a dynamic–thermodynamic snow–ice model with heat budgets for the upper and lower surface layers[21] and a viscous-plastic rheology[22]. BRIOS considers the ocean–ice-shelf interaction underneath ten Antarctic ice shelves[16,23] with time-invariant thicknesses, assuming the flux divergence to be in equilibrium with both the surface and the basal mass balance. The model has been successfully validated by the comparison with mooring and buoy observations regarding, for example, Weddell gyre transport[16], sea-ice thickness distribution and drift in the Weddell and Amundsen seas[24,25], and sea-ice concentration related to iceberg drift[26].

Ocean characteristics of the simulations forced with the output of both HadCM3-A and HadCM3-B for the twentieth century agree well with those from hindcasts using the NCEP-reanalysis[27]. In the following, we focus on the results of the runs forced with the output from HadCM3-B for the A1B scenario, because this scenario provides stronger signals and only HadCM3-B extends to the end of the twenty-second century, covering a period of 200 years. For the simulated present-day period, a slope front separates shelf water at the surface freezing point from relatively warm water that is advected to the southern Weddell Sea by the coastal current. However, starting in around 2036, pulses of warm water sporadically cross the 700-m-deep sill of the Filchner Trough at its eastern flank (Fig. 1) but do not reach the southern ice-shelf front (Fig. 2a). As early as 2070, water warmer than 0 °C begins to enter the Filchner Trough continuously (Fig. 2b), reaching the grounding lines of the southern tributaries 6 years later (Fig. 2c). After a further 14 years, the whole trough plus the southern half of the Ronne Ice Shelf cavity are filled with water of open-ocean origin (Fig. 2d). This corresponds to a warming of the deep southern cavity by more than 2 °C. The sporadic flow of warm water into the Filchner Trough during the twenty-first century, as well as its southward

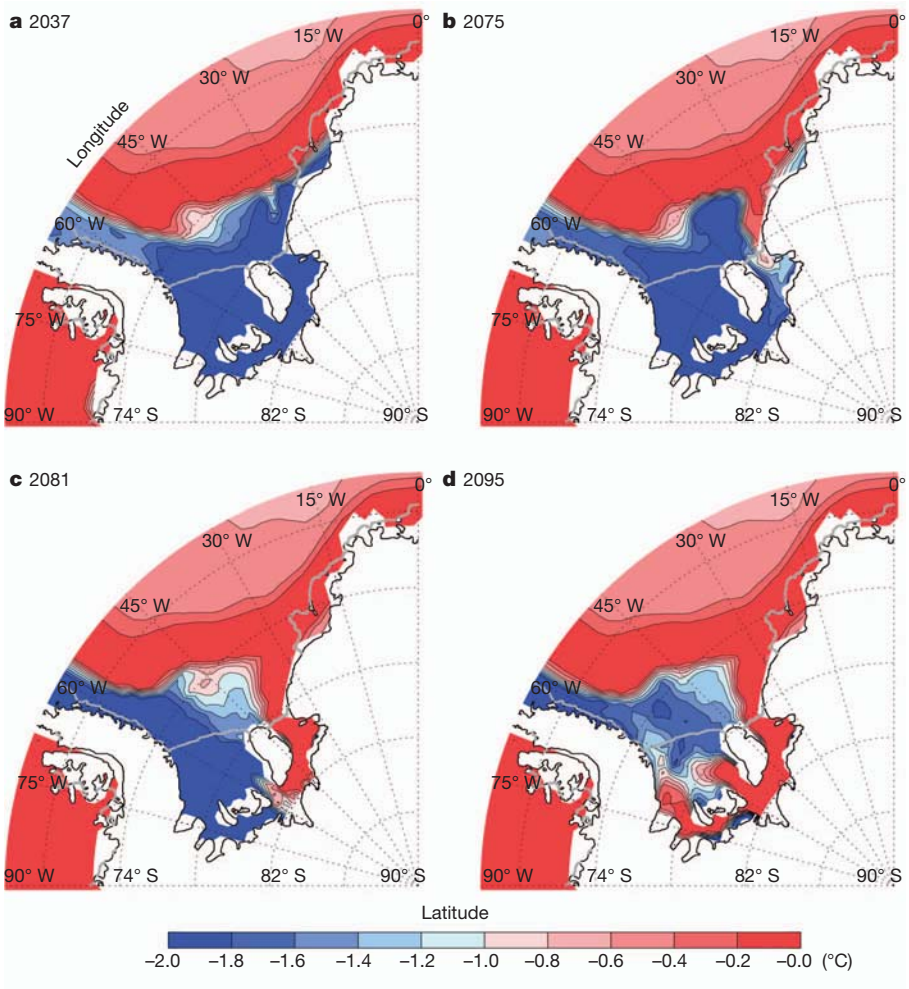


Figure 2 | Simulated evolution of near-bottom temperatures in the Weddell Sea. a–d, Values are from 60 m above bottom for the period 2030–2099 of the HadCM3-B/A1B scenario. Warm pulses into the Filchner Trough (2037; **a**) are followed by a return of the shelf water masses to the cold state typical for present conditions. The final (unrevoked) destruction of the slope front starts in 2066; by 2075 (**b**), the tongue of slightly modified warm deep water reaches the Filchner Ice Shelf front. It fills the deeper part of the Filchner Ice Shelf cavity and enters the Ronne Ice Shelf cavity near the grounding line south of Berkner Island in 2081 (**c**). By 2095 (**d**), warm water fills most of the bottom layer of the Filchner–Ronne Ice Shelf cavity, reaching a quasi-steady state. We note that a trend in the water mass properties of the interior Weddell Sea is not associated with any of these processes. The solid grey line off the coastline indicates the ice-shelf front.

propagation, is also suggested by results of the finite-element model FESOM[28] when forced with the HadCM3-B/A1B output (Supplementary Information). FESOM is a coupled ice–ocean model that also takes ice shelves into account, but it has a different architecture and a resolution that allows the simulation of eddies. Therefore, the model is expected to react more intensely to moderate perturbations in atmosphere and sea ice. Owing to the higher resolution of the marginal seas (~10 km) in FESOM, the warm water pulses reach the interior of the Filchner–Ronne Ice Shelf cavity less diluted (Supplementary Fig. 4) and thus cause earlier significant increases in basal mass loss (Supplementary Fig. 5).

The analysis of the forcing fields and the BRIOS output reveals that the redirection of the coastal current in the southeastern Weddell Sea is caused locally by an interplay between several climate components. During the twenty-first century, a continuous atmospheric surface warming (up to 4 °C per century) decreases the loss of sensible heat by the ocean. Together with an increase in long-wave downward radiation (up to 10 W m^{-2} per century) this reduces the thickness and concentration of the sea ice, allowing its drift speed to increase and, thus, a more efficient momentum transfer to the ocean surface off the Luitpold Coast (Fig. 3a, b). The enhanced surface stress, which is not related to an increase in atmospheric wind stress, directs the coastal current southwards towards the Filchner–Ronne Ice Shelf front, as it approaches the 700-m-deep sill of the Filchner Trough. The importance of the different atmospheric forcing variables to the redirection of the coastal current and, thus, the increase in melting at the base of the Filchner–Ronne Ice Shelf is investigated by means of additional sensitivity experiments (Supplementary Information). Because about 80% of the changes occur in the twenty-first century, these experiments are confined to the period 2000–2099. The first simulation applies detrended atmospheric forcing variables only, followed by runs in which the trends of 2-m temperature (the air temperature at an altitude of 2 m) and/or long-wave downward radiation were consecutively added.

The warming of the whole Filchner–Ronne Ice Shelf cavity by more than 2 °C boosts average basal melting from 0.2 m yr^{-1} to 4 m yr^{-1} at the end of the twenty-first century, with the maximum exceeding 50 m yr^{-1} near the deep southern grounding line. The values correspond to a jump in basal mass loss from 82 Gt yr^{-1} to ~1,600 Gt yr^{-1} (Fig. 3c), which represents 64% of the simulated circumpolar total. This total increases within two decades from ~1,000 Gt yr^{-1} to ~2,500 Gt yr^{-1}. In contrast, basal mass loss beneath the Ross Ice Shelf remains constant at ~80 Gt yr^{-1}. A similar drastic change in Filchner–Ronne Ice Shelf basal mass loss and circumpolar ice-shelf basal mass loss also happens in the simulations (Table 1) forced with the A1B output of HadCM3-A, but with a delay of 10 years, and the E1 output of HadCM3-B, but with a delay of 50 years, respectively (Fig. 3c). Owing to our assumption of fixed ice-shelf thicknesses, we cannot accurately predict basal mass losses for long periods of high melting. However, if we assume that grounding lines retreat into deeper basins[29], our melt rates have to be considered as lower bounds. In addition, numerical experiments show that ice shelves adjust to perturbations in ocean temperature on timescales ranging from several decades to a few centuries[30].

As a consequence of the increased input of fresh water due to ice-shelf basal melting, the Weddell Sea surface layer and the water masses on the whole southern and western continental shelves freshen rapidly. Today the high-salinity shelf water of these areas is one ingredient for the formation of deep and bottom waters of the Weddell Sea[7,31]. These water masses change their characteristics as the shelf water freshens.

Given the differences among the climate scenarios and the model realizations, we do not intend to predict the exact date of the changes in the circulation of the southern Weddell Sea. Instead, we emphasize the sensitivity of a small Antarctic coastal region to climate change with potentially severe consequences for the mass balance of a large Antarctic ice shelf. Determining the extent to which this influences the dynamics of the East Antarctic Ice Sheet will require further simulation, forcing a coupled ice-sheet–ice-shelf model with the predicted

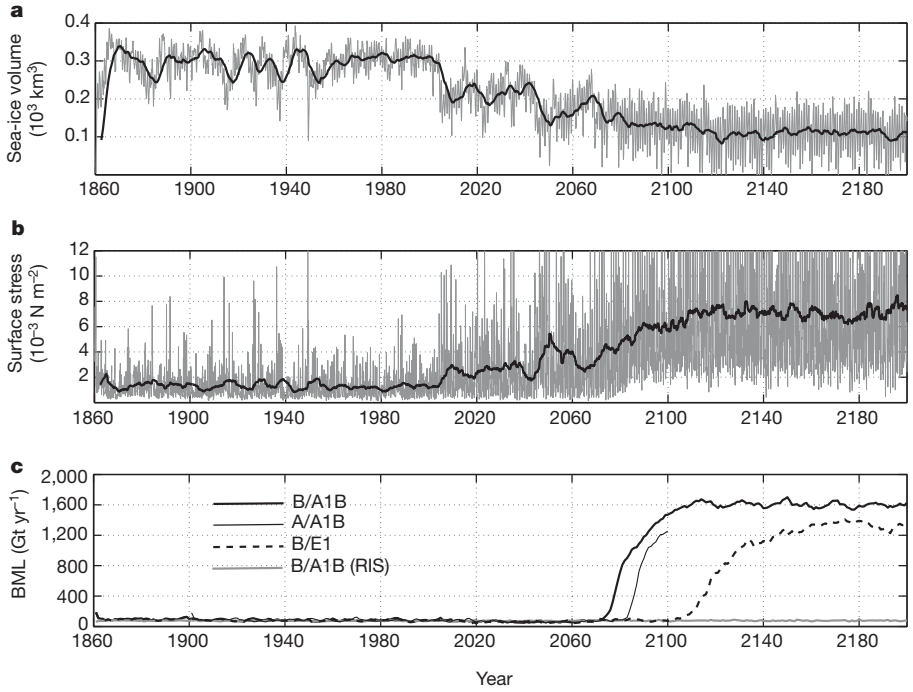


Figure 3 | Modelled time series (1860–2199) for the southeastern Weddell Sea. **a**, Area-integrated (Fig. 1) sea-ice volume for BRIOS forced with the twentieth-century and A1B atmospheric output of the climate model HadCM3-B. Grey and black lines represent monthly means and 5-year running means, respectively. **b**, Area-mean ocean-surface stress, for the same model as in **a**. Not only is the long-term decrease in the sea-ice volume reflected by an increase in the ocean-surface stress, but the coherence also holds for single events (for example, around 1940 and 2050). A correlation coefficient is not provided because of the dominance of the long-term variability. **c**, Basal mass losses (BMLs) in gigatonnes per year. Thin and thick lines represent simulations forced with the atmospheric output of the climate models HadCM3-A and HadCM3-B, respectively. HadCM3-A forcing is available only for the period 1900–2099 and the A1B scenario (Table 1). Solid and dashed lines represent results from forcing with twentieth-century and either A1B or, respectively, E1 output. Black lines show BML for the Filchner–Ronne Ice Shelf and the grey line shows that for the Ross Ice Shelf (RIS).

temperature perturbation. The use of the output of two different configurations of HadCM3 in different scenarios and the confirmation of the BRIOS results by FESOM, a coupled ice–ocean model with higher resolution and a different model architecture, reduces unavoidable uncertainties when dealing with processes related to climate change. Therefore, we are confident that our proposed mechanism is not a model artefact but quite a realistic mechanism. Consequently, we welcome the effort to monitor the coastal current during the upcoming expeditions to the southeastern Weddell Sea.

Received 7 July 2011; accepted 13 March 2012.

1. Walker, D. P. *et al.* Oceanic heat transport onto the Amundsen Sea shelf through a submarine glacial trough. *Geophys. Res. Lett.* **34,** L02602 (2007).
2. Hellmer, H. H., Jacobs, S. S. & Jenkins, A. in *Ocean, Ice, and Atmosphere: Interactions at the Antarctic Continental Margin* (eds Jacobs, S. S. & Weiss, R. F.) 83–99 (Antarctic Res. Ser. 75, American Geophysical Union, 1998).
3. Jacobs, S. S., Jenkins, A., Giulivi, C. & Dutrieux, P. Stronger ocean circulation and increased melting under Pine Island Glacier ice shelf. *Nature Geosci.* **4,** 519–523 (2011).
4. Payne, A. J. *et al.* Numerical modeling of ocean-ice interactions under Pine Island Bay's ice shelf. *J. Geophys. Res.* **112,** C10019 (2007).
5. Gordon, C. *et al.* The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dyn.* **16,** 147–168 (2000).
6. Rignot, E. *et al.* Acceleration of the contribution of the Greenland and Antarctic ice sheets to sea level. *Geophys. Res. Lett.* **38,** L05503 (2011).
7. Schröder, M. & Fahrbach, E. On the structure and the transport in the eastern Weddell Gyre. *Deep-Sea Res. II* **46,** 501–527 (1999).
8. Schröder, M., Hellmer, H. H. & Absy, J. M. On the near-bottom variability at the tip of the Antarctic Peninsula. *Deep-Sea Res. II* **49,** 4767–4790 (2002).
9. Nicholls, K. W., Boehme, L., Biuw, M. & Fedak, M. A. Wintertime ocean conditions over the southern Weddell Sea continental shelf, Antarctica. *Geophys. Res. Lett.* **35,** L21605 (2008).
10. Foldvik, A., Gammelsrød, T. & Tørresen, T. in *Oceanology of the Antarctic Continental Shelf* (ed. Jacobs, S. S.) 5–20 (Antarctic Res. Ser. 43, American Geophysical Union, 1985).
11. Makinson, K. & Nicholls, K. W. Modeling tidal currents beneath Filchner-Ronne Ice Shelf and the adjacent continental shelf: their effect on mixing and transport. *J. Geophys. Res.* **104,** 13449–13465 (1999).
12. Nicholls, K. W. Predicted reduction in basal melt rates of an Antarctic ice shelf in a warmer climate. *Nature* **388,** 460–462 (1997).
13. Bamber, J. L., Riva, R. E. M., Vermeersen, B. L. A. & LeBrocq, A. Reassessment of the potential sea-level rise from a collapse of the West Antarctic Ice Sheet. *Science* **324,** 901–903 (2009).
14. Jenkins, A. *et al.* Observations beneath Pine Island Glacier in West Antarctica and implications for its retreat. *Nature Geosci.* **3,** 468–472 (2010).
15. Bamber, J. L., Vaughan, D. G. & Joughin, I. Widespread complex flow in the interior of the Antarctic ice sheet. *Science* **287,** 1248–1250 (2000).
16. Beckmann, A., Hellmer, H. H. & Timmermann, R. A numerical model of the Weddell Sea: large-scale circulation and water mass distribution. *J. Geophys. Res.* **104,** 23375–23391 (1999).
17. Collins, M. *et al.* Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Clim. Dyn.* **36,** 1737–1766 (2011).
18. Johns, T. C. *et al.* Climate change under aggressive mitigation: The ENSEMBLES multi-model experiment. *Clim. Dyn.* **37,** 1975–2004 (2011).
19. Lowe, J. A. *et al.* New study for climate modelling, analyses, and scenarios. *Eos* **90,** 181–182 (2009).
20. Nakicevovic, N. *et al. IPCC Special Report on Emissions Scenarios* (Cambridge Univ. Press, 2000).
21. Parkinson, C. L. & Washington, W. M. A large-scale numerical model of sea ice. *J. Geophys. Res.* **84,** 311–337 (1979).
22. Hibler, W. D. III. A dynamic thermodynamic sea ice model. *J. Phys. Oceanogr.* **9,** 815–846 (1979).
23. Hellmer, H. H. Impact of Antarctic ice shelf basal melting on sea ice and deep ocean properties. *Geophys. Res. Lett.* **31,** L10307 (2004).
24. Timmermann, R., Beckmann, A. & Hellmer, H. H. Simulations of ice-ocean dynamics in the Weddell Sea: 1. Model configuration and validation. *J. Geophys. Res.* **107,** 3024 (2002).
25. Assmann, K. M., Hellmer, H. H. & Jacobs, S. S. Amundsen Sea ice production and transport. *J. Geophys. Res.* **110,** C12013 (2005).
26. Lichey, C. & Hellmer, H. H. Modeling giant-iceberg drift under the influence of sea ice in the Weddell Sea, Antarctica. *J. Glaciol.* **47,** 452–460 (2001).
27. Kalnay, E. M. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
28. Timmermann, R. *et al.* Ocean circulation and sea ice distribution in a finite element global sea ice–ocean model. *Ocean Model.* **27,** 114–129 (2009).
29. Timmermann, R. *et al.* A consistent dataset of Antarctic ice sheet topography, cavity geometry, and global bathymetry. *Earth Syst. Sci. Data* **2,** 261–273 (2010).
30. Walker, R. T. & Holland, D. M. A two-dimensional coupled model for ice shelf-ocean interaction. *Ocean Model.* **17,** 123–139 (2007).
31. Gordon, A. L., Visbeck, M. & Huber, B. Export of Weddell Sea deep and bottom water. *J. Geophys. Res.* **106,** 9005–9017 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Wübber and W. Cohrs for providing stable computer facilities at the Alfred-Wegener-Institute for Polar and Marine Research; the Ice2Sea community for discussions during project meetings; and J. Ridley, M. Martin and A. Levermann for comments on the manuscript. This work was supported by funding to the Ice2Sea programme from the European Union Seventh Framework Programme, grant number 226375. This is Ice2Sea contribution number 41.

Author Contributions H.H.H. had the idea to force BRIOS with Intergovernmental Panel on Climate Change scenarios, did 50% of the BRIOS simulations, conducted a significant part of the analysis of model output, wrote the main text of the paper and participated in the figure preparation. F.K. did 50% of the BRIOS simulations, conducted the analysis of the atmospheric forcing and wrote Supplementary Information. R.T. did all FESOM simulations, was involved in the analysis of model output and prepared most of the figures. J.D. provided the glaciological expertise for the interpretation of the model results related to basal mass loss. J.R. extracted the atmospheric forcings for all simulations and was involved in the analysis of model output. All authors participated in the discussion on model results and in drafting the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to H.H.H. (hartmut.hellmer@awi.de).

LETTER

doi:10.1038/nature11069

Comparing the yields of organic and conventional agriculture

Verena Seufert[1], Navin Ramankutty[1] & Jonathan A. Foley[2]

Numerous reports have emphasized the need for major changes in the global food system: agriculture must meet the twin challenge of feeding a growing population, with rising demand for meat and high-calorie diets, while simultaneously minimizing its global environmental impacts[1,2]. Organic farming—a system aimed at producing food with minimal harm to ecosystems, animals or humans—is often proposed as a solution[3,4]. However, critics argue that organic agriculture may have lower yields and would therefore need more land to produce the same amount of food as conventional farms, resulting in more widespread deforestation and biodiversity loss, and thus undermining the environmental benefits of organic practices[5]. Here we use a comprehensive meta-analysis to examine the relative yield performance of organic and conventional farming systems globally. Our analysis of available data shows that, overall, organic yields are typically lower than conventional yields. But these yield differences are highly contextual, depending on system and site characteristics, and range from 5% lower organic yields (rain-fed legumes and perennials on weak-acidic to weak-alkaline soils), 13% lower yields (when best organic practices are used), to 34% lower yields (when the conventional and organic systems are most comparable). Under certain conditions—that is, with good management practices, particular crop types and growing conditions—organic systems can thus nearly match conventional yields, whereas under others it at present cannot. To establish organic agriculture as an important tool in sustainable food production, the factors limiting organic yields need to be more fully understood, alongside assessments of the many social, environmental and economic benefits of organic farming systems.

Although yields are only part of a range of ecological, social and economic benefits delivered by farming systems, it is widely accepted that high yields are central to sustainable food security on a finite land basis[1,2]. Numerous individual studies have compared the yields of organic and conventional farms, but few have attempted to synthesize this information on a global scale. A first study of this kind[6] concluded that organic agriculture matched, or even exceeded, conventional yields, and could provide sufficient food on current agricultural land. However, this study was contested by a number of authors; the criticisms included their use of data from crops not truly under organic management and inappropriate yield comparisons[7,8].

We performed a comprehensive synthesis of the current scientific literature on organic-to-conventional yield comparisons using formal meta-analysis techniques. To address the criticisms of the previous study[6] we used several selection criteria: (1) we restricted our analysis to studies of 'truly' organic systems, defined as those with certified organic management or non-certified organic management, following the standards of organic certification bodies (see Supplementary Information); (2) we only included studies with comparable spatial and temporal scales for both organic and conventional systems (see Methods); and (3) we only included studies reporting (or from which we could estimate) sample size and error. Conventional systems were either high- or low-input commercial systems, or subsistence agriculture. Sixty-six studies met these criteria, representing 62 study sites, and reporting 316 organic-to-conventional yield comparisons on 34 different crop species (Supplementary Table 4).

The average organic-to-conventional yield ratio from our meta-analysis is 0.75 (with a 95% confidence interval of 0.71 to 0.79); that is, overall, organic yields are 25% lower than conventional (Fig. 1a). This result only changes slightly (to a yield ratio of 0.74) when the analysis is limited to studies following high scientific quality standards (Fig. 2). When comparing organic and conventional yields it is important

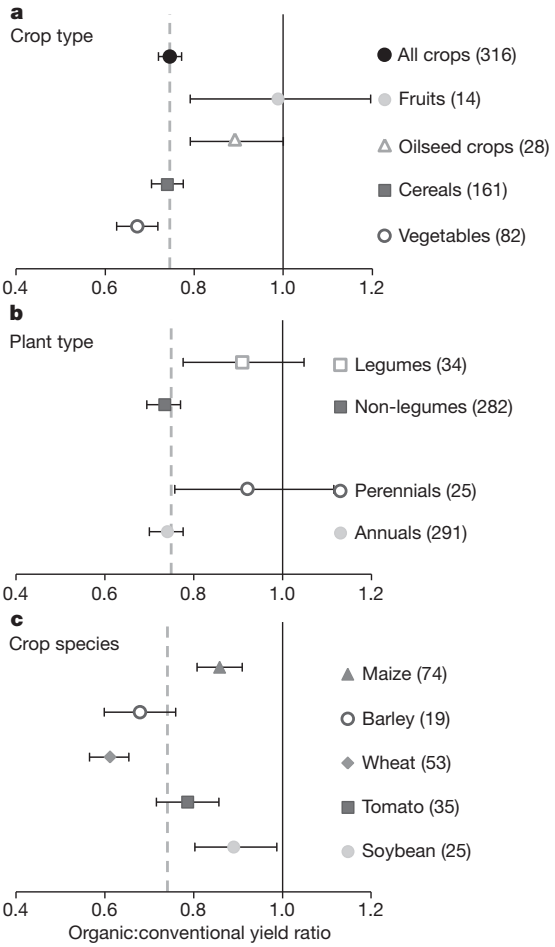


Figure 1 | Influence of different crop types, plant types and species on organic-to-conventional yield ratios. **a**–**c**, Influence of crop type (**a**), plant type (**b**) and crop species (**c**) on organic-to-conventional yield ratios. Only those crop types and crop species that were represented by at least ten observations and two studies are shown. Values are mean effect sizes with 95% confidence intervals. The number of observations in each class is shown in parentheses. The dotted line indicates the cumulative effect size across all classes.

[1]Department of Geography and Global Environmental and Climate Change Center, McGill University, Montreal, Quebec H2T 3A3, Canada. [2]Institute on the Environment (IonE), University of Minnesota, 1954 Buford Avenue, St Paul, Minnesota 55108, USA.

to consider the food output per unit area and time, as organic rotations often use more non-food crops like leguminous forage crops in their rotations[7]. However, the meta-analysis suggests that studies using longer periods of non-food crops in the organic rotation than conventional systems do not differ in their yield ratio from studies using similar periods of non-food crops (Fig. 2 and Supplementary Table 5). It thus appears that organic rotations do not require longer periods of non-food crops, which is also corroborated by the fact that the majority of studies (that is, 76%) use similar lengths of non-food crops in the organic and conventional systems.

The performance of organic systems varies substantially across crop types and species (Fig. 1a–c; see Supplementary Table 5 for details on categorical analysis). For example, yields of organic fruits and oilseed crops show a small (−3% and −11% respectively), but not statistically significant, difference to conventional crops, whereas organic cereals and vegetables have significantly lower yields than conventional crops (−26% and −33% respectively) (Fig. 1a).

These differences seem to be related to the better organic performance (referring to the relative yield of organic to conventional systems) of perennial over annual crops and legumes over non-legumes (Fig. 1b). However, note that although legumes and perennials (and fruits and oilseed crops) show statistically insignificant organic-to-conventional yield differences, this is owing to the large uncertainty range resulting from their relatively small sample size ($n = 34$ for legumes, $n = 25$ for perennials, $n = 14$ for fruits and $n = 28$ for oilseed crops; Fig. 1), and combining legumes and perennials reveals a significant, but small, yield difference (Fig. 2).

Part of these yield responses can be explained by differences in the amount of nitrogen (N) input received by the two systems (Fig. 3a). When organic systems receive higher quantities of N than conventional systems, organic performance improves, whereas conventional systems do not benefit from more N. In other words, organic systems appear to be N limited, whereas conventional systems are not. Indeed, N availability has been found to be a major yield-limiting factor in many organic systems[9]. The release of plant-available mineral N from organic sources such as cover crops, compost or animal manure is slow and often does not keep up with the high crop N demand during the peak growing period[9,10]. The better performance of organic legumes and perennials is not because they received more N, but rather because they seem to be more efficient at using N (Supplementary Table 7 and Supplementary Fig. 4). Legumes are not as dependent on external N sources as non-legumes, whereas perennials, owing to their longer growing period and extensive root systems, can achieve a better synchrony between nutrient demands and the slow release of N from organic matter[11].

Organic crops perform better on weak-acidic to weak-alkaline soils (that is, soils with a pH between 5.5. and 8.0; Fig. 3b). A possible explanation is the difficulty of managing phosphorus (P) in organic systems. Under strongly alkaline and acidic conditions, P is less readily available to plants as it forms insoluble phosphates, and crops depend to a stronger degree on soil amendments and fertilizers. Organic systems often do not receive adequate P inputs to replenish the P lost through harvest[12]. To test this hypothesis we need further research on the performance and nutrient dynamics of organic agriculture on soils of varying pH.

Studies that reported having applied best management practices in both systems show better organic performance (Fig. 3c). Nutrient and pest management in organic systems rely on biological processes to deliver plant nutrients and to control weed and herbivore populations. Organic yields thus depend more on knowledge and good management practices than conventional yields. However, in organic systems that are not N limited (as they grow perennial or leguminous crops, or apply large N inputs), best management practices are not required (Supplementary Table 11).

It is often reported that organic yields are low in the first years after conversion and gradually increase over time, owing to improvements in soil fertility and management skills[13]. This is supported by our analysis: organic performance improves in studies that lasted for more than two seasons or were conducted on plots that had been organic for at least 3 years (Fig. 2, Supplementary Fig. 5 and Supplementary Table 13).

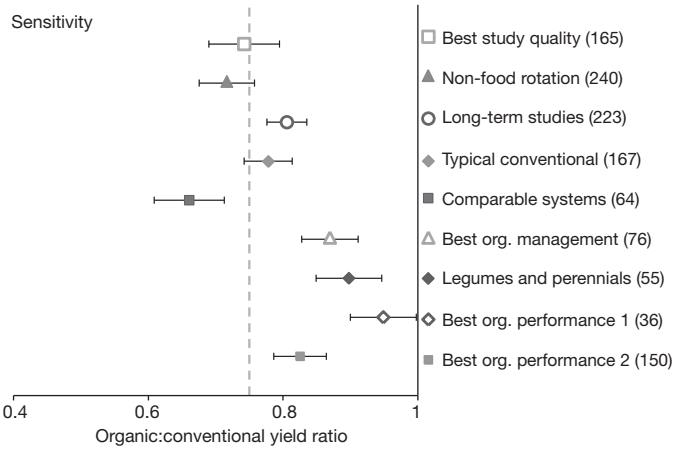


Figure 2 | Sensitivity study of organic-to-conventional yield ratios. Best study quality, peer-reviewed studies using appropriate study design and making appropriate inferences; non-food rotation, studies where both systems have a similar duration of non-food crops; long-term studies, excludes very short duration and recently converted studies; typical conventional, restricted to commercial conventional systems with yields comparable to local averages; comparable systems, studies that use appropriate study design and make appropriate inferences, where both systems have the same non-food rotation length and similar N inputs; best org. management, excludes studies without best management practices or crop rotations; legumes and perennials, restricted to leguminous and perennial crops; best org. performance 1, rain-fed legumes and perennials on weak-acidic to weak-alkaline soils; best org. performance 2, rain-fed and weak-acidic to weak-alkaline soils. Values are mean effect sizes with 95% confidence intervals. The number of observations is shown in parentheses. The dotted line indicates the effect size across all studies.

Water relations also influence organic yield ratios—organic performance is −35% under irrigated conditions, but only −17% under rain-fed conditions (Fig. 3e). This could be due to a relatively better organic performance under variable moisture conditions in rain-fed systems. Soils managed with organic methods have shown better water-holding capacity and water infiltration rates and have produced higher yields than conventional systems under drought conditions and excessive rainfall[14,15] (see Supplementary Information). On the other hand, organic systems are often nutrient limited (see earlier), and thus probably do not respond as strongly to irrigation as conventional systems.

The majority of studies in our meta-analysis come from developed countries (Supplementary Fig. 1). Comparing organic agriculture across the world, we find that in developed countries organic performance is, on average, −20%, whereas in developing countries it is −43% (Fig. 3f). This poor performance of organic agriculture in developing countries may be explained by the fact that a majority of the data (58 of 67 observations) from developing countries seem to have atypical conventional yields (>50% higher than local yield averages), coming from irrigated lands (52 of 67), experimental stations (54 of 67) and from systems not using best management practices (67 of 67; Supplementary Fig. 10 and Supplementary Table 8). In the few cases from developing countries where organic yields are compared to conventional yields typical for the location or where the yield data comes from surveys, organic yields do not differ significantly from conventional yields because of a wide confidence interval resulting from the small sample size ($n = 8$ and $n = 12$ respectively, Supplementary Fig. 10a).

The results of our meta-analysis differ dramatically from previous results[6]. Although our organic performance estimate is lower than previously reported[6] in developed countries (−20% compared to −8%), our results are markedly different in developing countries

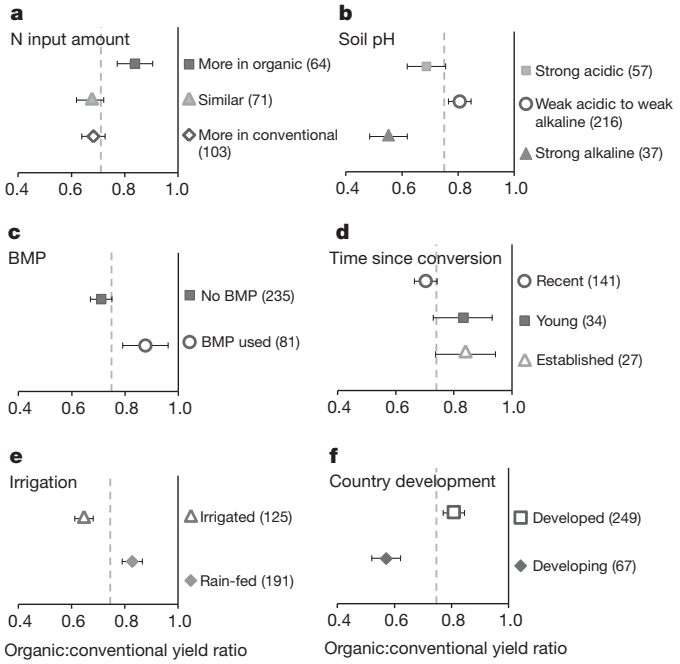


Figure 3 | Influence of N input, soil pH, best management practices, time since conversion to organic management, irrigation and country development. a–f, Influence of the amount of N input (**a**), soil pH (**b**), the use of best management practices (BMP; **c**), time since conversion to organic management (**d**), irrigation (**e**) and country development (**f**) on organic-to-conventional yield ratios. For details on the definition of categorical variables see Supplementary Tables 1–3. Values are mean effect sizes with 95% confidence intervals. The number of observations in each class is shown in parentheses. The dotted line indicates the cumulative effect size across all classes.

(−43% compared to +80%). This is because the previous analysis mainly included yield comparisons from conventional low-input subsistence systems, whereas our data set mainly includes data from high-input systems for developing countries. However, the previous study compared subsistence systems to yields that were not truly organic, and/or from surveys of projects that lacked an adequate control. Not a single study comparing organic to subsistence systems met our selection criteria and could be included in the meta-analysis. We cannot, therefore, rule out the claim[16] that organic agriculture can increase yields in smallholder agriculture in developing countries. But owing to a lack of quantitative studies with appropriate controls we do not have sufficient scientific evidence to support it either. Fortunately, the Swiss Research Institute of Organic Agriculture (FiBL) recently established the first long-term comparison of organic and different conventional systems in the tropics[17]. Such well-designed long-term field trials are urgently needed.

Our analysis shows that yield differences between organic and conventional agriculture do exist, but that they are highly contextual. When using best organic management practices yields are closer to (−13%) conventional yields (Fig. 2). Organic agriculture also performs better under certain agroecological conditions—for example, organic legumes or perennials, on weak-acidic to weak-alkaline soils, in rain-fed conditions, achieve yields that are only 5% lower than conventional yields (Fig. 2). On the other hand, when only the most comparable conventional and organic systems are considered the yield difference is as high as 34% (Fig. 2). In developed countries or in studies that use conventional yields that are representative of regional averages, the yield difference between comparable organic and conventional systems, however, goes down to 8% and 13%, respectively (see Supplementary Information).

In short, these results suggest that today's organic systems may nearly rival conventional yields in some cases—with particular crop types, growing conditions and management practices—but often they do not. Improvements in management techniques that address factors limiting yields in organic systems and/or the adoption of organic agriculture under those agroecological conditions where it performs best may be able to close the gap between organic and conventional yields.

Although we were able to identify some factors contributing to variations in organic performance, several other potentially important factors could not be tested owing to a lack of appropriate studies. For example, we were unable to analyse tillage, crop residue or pest management. Also, most studies included in our analysis experienced favourable growing conditions (Supplementary Fig. 8), and organic systems were mostly compared to commercial high-input systems (which had predominantly above-average yields in developing countries; Supplementary Figs 6b and 10a). In addition, it would be desirable to examine the total human-edible calorie or net energy yield of the entire farm system rather than the biomass yield of a single crop species. To understand better the performance of organic agriculture, we should: (1) systematically analyse the long-term performance of organic agriculture under different management regimes; (2) study organic systems under a wider range of biophysical conditions; (3) examine the relative yield performance of smallholder agricultural systems; and (4) evaluate the performance of farming systems through more holistic system metrics.

As emphasized earlier, yields are only part of a range of economic, social and environmental factors that should be considered when gauging the benefits of different farming systems. In developed countries, the central question is whether the environmental benefits of organic crop production would offset the costs of lower yields (such as increased food prices and reduced food exports). Although several studies have suggested that organic agriculture can have a reduced environmental impact compared to conventional agriculture[18,19], the environmental performance of organic agriculture per unit output or per unit input may not always be advantageous[20,21]. In developing countries, a key question is whether organic agriculture can help alleviate poverty for small farmers and increase food security. On the one hand, it has been suggested that organic agriculture may improve farmer livelihoods owing to cheaper inputs, higher and more stable prices, and risk diversification[16]. On the other hand, organic agriculture in developing countries is often an export-oriented system tied to a certification process by international bodies, and its profitability can vary between locations and years[22,23].

There are many factors to consider in balancing the benefits of organic and conventional agriculture, and there are no simple ways to determine a clear 'winner' for all possible farming situations. However, instead of continuing the ideologically charged 'organic versus conventional' debate, we should systematically evaluate the costs and benefits of different management options. In the end, to achieve sustainable food security we will probably need many different techniques—including organic, conventional, and possible 'hybrid' systems[24]—to produce more food at affordable prices, ensure livelihoods for farmers, and reduce the environmental costs of agriculture.

METHODS SUMMARY

We conducted a comprehensive literature search, compiling scientific studies comparing organic to conventional yields that met our selection criteria. We minimized the use of selection criteria based on judgments of study quality but examined its influence in the categorical analysis. We collected information on several study characteristics reported in the papers and derived characteristics of the study site from spatial global data sets (see Supplementary Tables 1–3 for a description of all categorical variables). We examined the difference between organic and conventional yields with the natural logarithm of the response ratio (the ratio between organic and conventional yields), an effect size commonly used in meta-analyses[25]. To calculate the cumulative effect size we weighted each individual observation by the inverse of the mixed-model variance. Such a categorical meta-analysis should be used when the data have some underlying structure and individual observations can be categorized into groups (for example, crop species or fertilization practices)[26]. An effect size is considered significant if its confidence interval does not overlap with 1 in the back-transformed response ratio. To test the influence of categorical variables on yield effect sizes we examined between-group

heterogeneity (Q_B). A significant Q_B indicates that there are differences in effect sizes between different classes of a categorical variable[26]. All statistical analyses were carried out in MetaWin 2.0[26].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 6 November 2011; accepted 9 March 2012.
Published online 25 April 2012.

1. Godfray, H. C. *et al.* Food security: the challenge of feeding 9 billion people. *Science* **327,** 812–818 (2010).
2. Foley, J. *et al.* Solutions for a cultivated planet. *Nature* **478,** 337–342 (2011).
3. McIntyre, B. D., Herren, H. R., Wakhungu, J. & Watson, R. T. *International Assessment of Agricultural Knowledge, Science and Technology for Development: Global Report* http://www.agassessment.org/ (Island, 2009).
4. De Schutter, O. *Report Submitted by the Special Rapporteur on the Right to Food* http://www2.ohchr.org/ english/issues/food/docs/A-HRC-16–49.pdf (United Nations, 2010).
5. Trewavas, A. Urban myths of organic farming. *Nature* **410,** 409–410 (2001).
6. Badgley, C. *et al.* Organic agriculture and the global food supply. *Renew. Agr. Food Syst.* **22,** 86–108 (2007).
7. Cassman, K. G. Editorial response by Kenneth Cassman: can organic agriculture feed the world-science to the rescue? *Renew. Agr. Food Syst.* **22,** 83–84 (2007).
8. Connor, D. J. Organic agriculture cannot feed the world. *Field Crops Res.* **106,** 187–190 (2008).
9. Berry, P. *et al.* Is the productivity of organic farms restricted by the supply of available nitrogen? *Soil Use Manage.* **18,** 248–255 (2002).
10. Pang, X. & Letey, J. Organic farming: challenge of timing nitrogen availability to crop nitrogen requirements. *Soil Sci. Soc. Am. J.* **64,** 247–253 (2000).
11. Crews, T. E. & Peoples, M. B. Can the synchrony of nitrogen supply and crop demand be improved in legume and fertilizer-based agroecosystems? A review. *Nutr. Cycl. Agroecosyst.* **72,** 101–120 (2005).
12. Oehl, F. *et al.* Phosphorus budget and phosphorus availability in soils under organic and conventional farming. *Nutr. Cycl. Agroecosyst.* **62,** 25–35 (2002).
13. Martini, E., Buyer, J. S., Bryant, D. C., Hartz, T. K. & Denison, R. F. Yield increases during the organic transition: improving soil quality or increasing experience? *Field Crops Res.* **86,** 255–266 (2004).
14. Letter, D., Seidel, R. & Liebhardt, W. The performance of organic and conventional cropping systems in an extreme climate year. *Am. J. Altern. Agric.* **18,** 146–154 (2003).
15. Colla, G. *et al.* Soil physical properties and tomato yield and quality in alternative cropping systems. *Agron. J.* **92,** 924–932 (2000).
16. Scialabba, N. & Hattam, C. *Organic Agriculture, Environment and Food Security* (Food and Agriculture Organization, 2002).
17. Research Institute of Organic Agriculture (FiBL). *Farming System Comparison in the Tropics* http://www.systems-comparison.fibl.org/ (2011).
18. Crowder, D. W., Northfield, T. D., Strand, M. R. & Snyder, W. E. Organic agriculture promotes evenness and natural pest control. *Nature* **466,** 109–112 (2010).
19. Bengtsson, J., Ahnström, J. & Weibull, A.-C. The effects of organic agriculture on biodiversity and abundance: a meta-analysis. *J. Appl. Ecol.* **42,** 261–269 (2005).
20. Kirchmann, H. & Bergström, L. Do organic farming practices reduce nitrate leaching? *Commun. Soil Sci. Plan.* **32,** 997–1028 (2001).
21. Leifeld, J. & Fuhrer, J. Organic farming and soil carbon sequestration: what do we really know about the benefits? *Ambio* **39,** 585–599 (2010).
22. Valkila, J. Fair trade organic coffee production in Nicaragua—sustainable development or a poverty trap? *Ecol. Econ.* **68,** 3018–3025 (2009).
23. Raynolds, L. T. The globalization of organic agro-food networks. *World Dev.* **32,** 725–743 (2004).
24. National Research Council. *Toward Sustainable Agricultural Systems in the 21st Century* (National Academies, 2010).
25. Hedges, L. V., Gurevitch, J. & Curtis, P. S. The meta-analysis of response ratios in experimental ecology. *Ecology* **80,** 1150–1156 (1999).
26. Rosenberg, M. S., Gurevitch, J. & Adams, D. C. *MetaWin: Statistical Software for Meta-analysis: Version 2* (Sinauer, 2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to the authors of the 66 studies whose extensive field work provided the data for this meta-analysis. Owing to space limitations our citations can be found in Supplementary Material. We would like to thank J. Reganold for useful comments on our manuscript. We are grateful to I. Perfecto, T. Moore, C. Halpenny, G. Seufert and S. Lehringer for valuable discussion and/or feedback on the manuscript and L. Gunst for sharing publications on the FiBL trials. D. Plouffe helped with the figures and M. Henry with compiling data. This research was supported by a Discovery Grant awarded to N.R. from the Natural Science and Engineering Research Council of Canada.

Author Contributions V.S. and N.R. designed the study. V.S. compiled the data and carried out data analysis. All authors discussed the results and contributed to writing the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to V.S. (verena.seufert@mail.mcgill.ca).

METHODS

Literature search. We searched the literature on studies reporting organic-to-conventional yield comparisons. First we used the references included in the previous study[6] and then extended the search by using online search engines (Google scholar, ISI web of knowledge) as well as reference lists of published articles. We applied several selection criteria to address the criticisms of the previous study[6] and to ensure that minimum scientific standards were met. Studies were only included if they (1) reported yield data on individual crop species in an organic treatment and a conventional treatment, (2) the organic treatment was truly organic (that is, either certified organic or following organic standards), (3) reported primary data, (4) the scale of the organic and conventional yield observations were comparable, (5) data were not already included from another paper (that is, avoid multiple counting), and (6) reported the mean (X), an error term (standard deviation (s.d.), standard error (s.e.) or confidence interval) and sample size (n) as numerical or graphical data, or if X and s.d. of yields over time could be calculated from the reported data. For organic and conventional treatments to be considered comparable, the temporal and spatial scale of the reported yields needed to be the same, that is, national averages of conventional agriculture compared to national averages of organic agriculture or yields on an organic farm compared to yields on a neighbouring conventional farm—not included were, for example, single farm yields compared to national or regional averages or before–after comparisons. Previous studies[27] have illustrated the danger of comparing yield data drawn from single plots and field trials to larger state and national averages.

The use of selection criteria is a critical step in conducting a meta-analysis. On the one hand, scientific quality and comparability of observations needs to be ensured. On the other hand, a meta-analysis should provide as complete a summary of the current research as possible. There is an ongoing debate about whether meta-analyses should adopt very specific selection criteria to prevent mixing incomparable data sets together and to minimize variation in the data set[28] or whether, instead, meta-analyses should include as wide a range of studies as possible to allow for an analysis of sources of variation[29]. We followed the generally recommended approach, trying to minimize the use of selection criteria based on judgments of study quality[30]. Instead, we examined the influence of quality criteria empirically by evaluating the differences between observations with different quality standards. We did not therefore exclude yield observations from non-peer-reviewed sources or from studies that lacked an appropriate experimental design a priori. The quality of the study and the comparability of the organic and conventional systems were assessed by evaluating the experimental design of the study as well as the form of publication. Studies that were published in peer-reviewed journals and that controlled for the possible influence of variability in space and time on experimental outcomes through an appropriate experimental design were considered to follow high quality standards.

Categorical variables. In addition to study quality criteria, information on several other study characteristics like crop species, location and timescale, and on different management practices, was collected (see Supplementary Tables 1–3). We also wanted to test the effect of study site characteristics on yield ratios and we thus collected information on biophysical characteristics of the study site. As most studies did not report climate or soil variables we derived information on several agroecological variables that capture cropland suitability[31], including the moisture index α (the ratio of actual to potential evapotranspiration) as an indicator of moisture availability to crops, growing degree days (GDD, the annual sum of daily mean temperatures over a base temperature of 5 °C) as an indicator of growing season length, as well as soil carbon density (C_{soil}, as a measure of soil organic content) and soil pH as indicators of soil quality from the latitude × longitude values of the study site and global spatial models/data sets at 5 min resolution[32,33].

We derived the thresholds for the classification of these climate and soil variables from the probability of cultivation functions previously described[31]. This probability of cultivation function is a curve fitted to the empirical relationship between cropland areas, α, GDD or C_{soil}. It describes the probability that a location with a certain climate or soil characteristic is covered by cropland. Suitable locations with favourable climate and soil characteristics have a higher probability of being cultivated. Favourable climate and soil characteristics can thus be inferred from the probability of cultivation. For α, GDD and C_{soil} a probability of cultivation under 30% was classified as 'low' suitability, between 30% and 70% as 'medium' suitability, and above 70% as 'high' suitability (Supplementary Table 3). Sites with low and medium suitable moisture indices are interpreted as having insufficient water availability, sites with low and medium GDD have short growing seasons, and sites with low and medium soil carbon densities are either unfertile because they have too small a C_{soil} and low organic matter content (and thus insufficient nutrients) or too high a C_{soil} in soils in wetlands where organic matter accumulates because they are submerged under water. For soil pH, instead, we defined thresholds based on expert judgment. Soil pH information was often given in the studies and we only derived soil pH values from the global data set if no soil pH value was indicated in the paper.

To assess whether the conventional yield values reported by studies and included in the meta-analysis were representative of regional average crop yields, we compared them to FAOSTAT yield data and a high-resolution spatial yield data set[34,35]. We used the FAO data[35], which reports national yearly crop yields from 1961 to 2009, for temporal detail and a yield data set[34], which reports sub-national crop yields for 175 crops for the year 2000 at a 5-min latitude by 5-min longitude resolution, for spatial detail. We calculated country average crop yields from FAO data for the respective study period and calculated the ratio of this average study-period yield to the year-2000 FAO national yield value. We derived the year-2000 yield value from the spatial data set through the latitude by longitude value of the study site and scaled this value to the study-period-to-year-2000 ratio from FAOSTAT. If the meta-analysis conventional yield value was more than 50% higher than the local yield average derived by this method it was classified as 'above average', when it was more than 50% lower as 'below average', and when it was within ±50% of local yield averages as 'comparable'. We choose this large yield difference as a threshold to account for uncertainties in the FAOSTAT and global yield data set[34].

Meta-analysis. The natural log of the response ratio[25] was used as an effect size metric for the meta-analysis. The response ratio is calculated as the ratio between the organic and the conventional yield. The use of the natural logarithm linearizes the metric (treating deviations in the numerator and the denominator the same) and provides more normal sampling distribution in small samples[25]. If the data set has some underlying structure and studies can be categorized into more than one group (for example, different crop species, or different fertilizer types) a categorical meta-analysis can be conducted[26]. Observations with the same or similar management or system characteristics were grouped together. We then used a mixed effects model to partition the variance of the sample, assuming that there is random variation within a group and fixed variation between groups. We calculated a cumulative effect size as weighted mean from all studies by weighting each individual observation by the reciprocal of the mixed-model variance, which is the sum of the study sampling variance and the pooled within-group variance. Weighted parametric meta-analysis should be used whenever possible to deal with heteroscedasticity in the sample and to increase the statistical power of the analysis[36]. The cumulative effect size is considered to be significantly different from zero (that is, the organic treatment shows a significant effect on crop yield) if its 95% confidence interval does not overlap zero.

To test for differences in the effect sizes between groups the total heterogeneity of the sample was partitioned into the within group (Q_W) and between group heterogeneity (Q_B) in a process similar to an analysis of variance[37]. The significance of Q_B was tested by comparing it against the critical value of the χ^2 distribution. A significant Q_B implies that there are differences among cumulative effect sizes between groups[26,38]. Only those effects that showed a significant Q_B are presented in graphs. All statistical analyses were carried out using MetaWin 2.0[26]. For representation in graphs effect sizes were back-transformed to response ratios.

Each observation in a meta-analysis is required to be independent. Repeated measurements in the same location over time are not independent. If yield values from a single experiment were reported over several years therefore the average yield over time was calculated and used in the meta-analysis. If the mean and variance of multiple years was reported, the weighted average over time was calculated by weighting each year by the inverse of its variance. Different experiments (for example, different tillage practices, crop species or fertilizer rates) from the same study are not necessarily independent. However, it is recommended to still include different experiments from the same study, as their omission would cause more distortions of the results than the lack of true independence[38]. We therefore included different experiments from a single study separately in the meta-analysis.

If data from the same experiment from the same study period were reported in several papers, the data were only included once, namely from the paper that reported the data in the highest detail (that is, reporting s.e./s.e. and n and/or reporting the longest time period). If instead data from the same experiment from different years were reported in separate papers, the data were included separately in the analysis (for example, refs 39, 40).

In addition to potential within-study dependence of effect size data, there can also be issues with between-study dependence of data[36]—data from studies conducted by the same author, in the same location or on the same crop species are also potentially non-independent. We addressed this issue by conducting a hierarchical, categorical meta-analysis (as described earlier), specifically testing for the influence of numerous moderators on the effect size. In addition, we examined the interaction between categorical variables through a combination of contingency

tables and sub-categorical analysis (see Supplementary Information for the results of this analysis and for a more detailed discussion of this issue).

We performed a sensitivity analysis (see Supplementary Table 14) to compare the robustness of results under more strict quality criteria (see discussion of definition of study quality earlier) and to assess organic yield ratios under a couple of specific system comparisons.

27. Johnston, M., Foley, J. A., Holloway, T., Kucharik, C. & Monfreda, C. Resetting global expectations from agricultural biofuels. *Environ. Res. Lett.* **4,** 014004 (2009).
28. Whittaker, R. J. Meta-analyses and mega-mistakes: calling time on meta-analysis of the species richness–productivity relationship. *Ecology* **91,** 2522–2533 (2010).
29. Hillebrand, H. & Cardinale, B. J. A critique for meta-analyses and the productivity–diversity relationship. *Ecology* **91,** 2545–2549 (2010).
30. Englund, G., Sarnelle, O. & Cooper, S. D. The importance of data-selection criteria: meta-analyses of stream predation experiments. *Ecology* **80,** 1132–1141 (1999).
31. Ramankutty, N., Foley, J. A., Norman, J. & McSweeney, K. The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Glob. Ecol. Biogeogr.* **11,** 377–392 (2002).
32. Deryng, D., Sacks, W., Barford, C. & Ramankutty, N. Simulating the effects of climate and agricultural management practices on global crop yield. *Glob. Biogeochem. Cycles* **25,** GB2006 (2011).
33. IGBP-DIS. *Soildata (V 0): A Program for Creating Global Soil-Property Databases* (IGBP Global Soils Data Task, 1998).
34. Monfreda, C., Ramankutty, N. & Foley, J. A. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Glob. Biogeochem. Cycles* **22,** GB1022 (2008).
35. Food and Agriculture Organization of the United Nations (FAO). *FAOSTAT* http://faostat.fao.org (2011).
36. Gurevitch, J. & Hedges, L. V. Statistical issues in ecological meta-analyses. *Ecology* **80,** 1142–1149 (1999).
37. Hedges, L. V. & Olkin, I. *Statistical Methods for Meta-Analysis*. (Academic, 1985).
38. Gurevitch, J., Morrow, L. L., Wallace, A. & Walsh, J. S. A meta-analysis of competition in field experiments. *Am. Nat.* **140,** 539–572 (1992).
39. Liebhardt, W. *et al.* Crop production during conversion from conventional to low-input methods. *Agron. J.* **81,** 150–159 (1989).
40. Drinkwater, L., Janke, R. & Rossoni-Longnecker, L. Effects of tillage intensity on nitrogen dynamics and productivity in legume-based grain systems. *Plant Soil* **227,** 99–113 (2000).

LETTER

doi:10.1038/nature11020

Selective cortical representation of attended speaker in multi-talker speech perception

Nima Mesgarani[1] & Edward F. Chang[1]

Humans possess a remarkable ability to attend to a single speaker's voice in a multi-talker background[1–3]. How the auditory system manages to extract intelligible speech under such acoustically complex and adverse listening conditions is not known, and, indeed, it is not clear how attended speech is internally represented[4,5]. Here, using multi-electrode surface recordings from the cortex of subjects engaged in a listening task with two simultaneous speakers, we demonstrate that population responses in non-primary human auditory cortex encode critical features of attended speech: speech spectrograms reconstructed based on cortical responses to the mixture of speakers reveal the salient spectral and temporal features of the attended speaker, as if subjects were listening to that speaker alone. A simple classifier trained solely on examples of single speakers can decode both attended words and speaker identity. We find that task performance is well predicted by a rapid increase in attention-modulated neural selectivity across both single-electrode and population-level cortical responses. These findings demonstrate that the cortical representation of speech does not merely reflect the external acoustic environment, but instead gives rise to the perceptual aspects relevant for the listener's intended goal.

Separating out a speaker of interest from other speakers in a noisy, crowded environment is a perceptual feat that we perform routinely. The ease with which we hear under these conditions belies the intrinsic complexity of this process, known as the cocktail party problem[1–3,6]: concurrent complex sounds, which are completely mixed upon entering the ear, are re-segregated and selected from within the auditory system. The resulting percept is that we selectively attend to the desired speaker while tuning out the others.

Although previous studies have described neural correlates of masking and selective attention to speech[4,5,7–9], fundamental questions remain unanswered regarding the precise nature of speech representation at the juncture where competing signals are resolved. In particular, when attending to a speaker within a mixture, it is unclear what key aspects (for example, spectrotemporal profile, spoken words and speaker identity) are represented in the auditory system and how they compare to representations of that speaker alone; how rapidly a selective neural representation builds up when one attends to a specific speaker; and whether breakdowns in these processes can explain distinct perceptual failures, such as the inability to hear the correct words, or follow the intended speaker.

To answer these questions, we recorded cortical activity from human subjects implanted with customized high-density multi-electrode arrays as part of their clinical work-up for epilepsy surgery[10]. Although limited to this clinical setting, these recordings provide simultaneous high spatial and temporal resolution while sampling the population neural activity from the non-primary auditory speech cortex in the posterior superior temporal lobe. We focused our analysis on high gamma (75–150 Hz) local field potentials[11], which have been found to correlate well with the tuning of multi-unit spike recordings[12]. In humans, the posterior superior temporal gyrus has been heavily implicated in speech perception[13], and is anatomically defined as the lateral parabelt auditory cortex (including Brodmann areas 41, 42 and 22)[14].

Subjects listened to speech samples from a corpus commonly used in multi-talker communication research[15,16]. A typical sentence was "ready tiger go to red two now" where "tiger" is the call sign, and "red two" is the colour–number combination. One male and one female speaker were selected, each speaking the same 12 unique combinations of two call signs (ringo or tiger), three colours (red, blue or green) and three numbers (two, five or seven). Example acoustic spectrograms from two individual speakers are shown in Fig. 1a, b. The two voices differ along several dimensions including pitch (male versus female), spectral profile (different vocal track shapes) and temporal characteristics (speaking rate). Subjects first listened to each of the speakers alone and were able to report the colour and number with 100% accuracy. Subjects then listened to a monaural, simultaneous mixture of the two speakers' phrases with different call signs, colours and numbers. The subjects were instructed to respond by indicating the colour and number spoken by the talker who uttered the target call sign. The target call sign (ringo or tiger) was fixed and shown visually on a monitor during each trial block, which contained 28 different mixture sounds. As the target speaker was changed randomly from trial to trial, the subjects were required to monitor both voices initially (divided attention) to identify the target speaker. The target call sign was switched after each block, turning the previous target speaker in each mixture into a masker. This resulted in two sets of behavioural and neural responses for each identical mixture sound, which differed only in the focus of attention. Subjects reported correct responses in 74.8% of trials.

Figure 1c illustrates the mixture spectrogram and how difficult it is to tell which sound parts belong to one speaker versus the other. The energy for both speakers is distributed broadly across the spectral and temporal domains, with overlap in some areas and isolated sound parts in others, as shown in their difference spectrogram (Fig. 1d; average spectrograms in Supplementary Fig. 1a).

To determine the spectrotemporal encoding of the attended speaker, the method of stimulus reconstruction was used[17–19] to estimate the speech spectrogram represented by the population neural responses. Reconstructed spectrograms provide an intuitive way to examine how the population neural responses encode the spectrotemporal features of speech, and more importantly, can be compared with the original acoustic spectrograms as well as across attentional conditions. We first calculated the reconstruction filters from a passive listening task using a separate continuous speech corpus (TIMIT[20]) that consisted of 499 unique short sentences spoken by 402 different speakers. The filters were then fixed and applied to a novel set of population neural responses to the single and attended mixture speech for spectrogram reconstruction.

When listening to a single speaker alone, the reconstructed spectrograms from population neural activity corresponded well to the spectrotemporal features of the original acoustic spectrograms (Fig. 1e, f compared to Fig. 1a, b, respectively), exhibiting fairly precise temporal features and spectral selectivity (for example, correspondence between the high frequency bursts of energy in "tiger" and "two", in Fig. 1a, b, e, f).

[1]Departments of Neurological Surgery and Physiology, UCSF Center for Integrative Neuroscience, University of California, San Francisco, California 94143, USA.

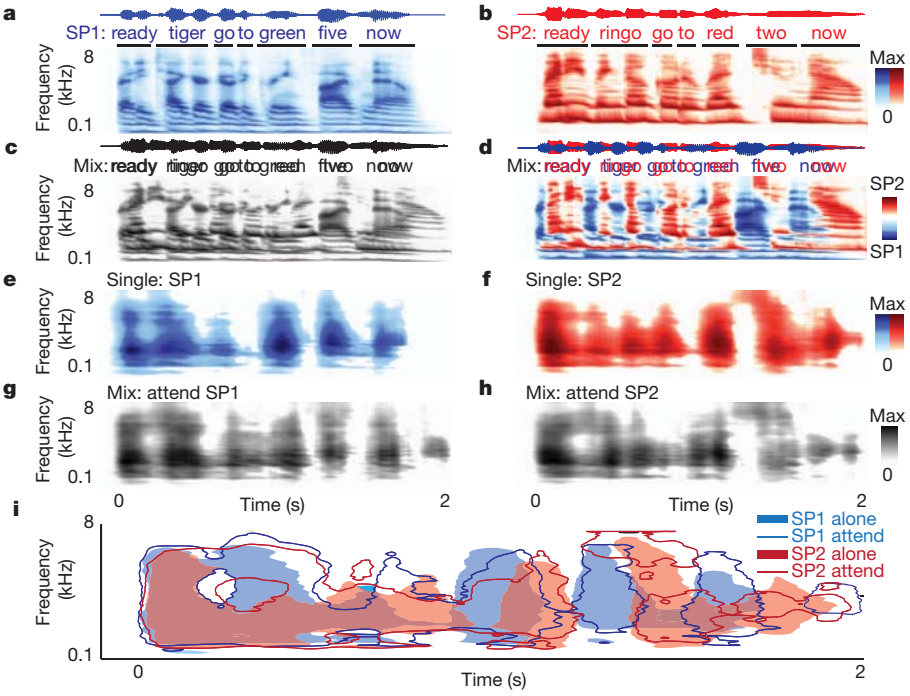


Figure 1 | Acoustic and neural reconstructed spectrograms for speech from a single speaker or a mixture of speakers. **a**, **b**, Example acoustic waveform and auditory spectrograms of speaker one (male; **a**) and speaker two (female; **b**). **c**, Waveform and spectrogram of the mixture of the two shows highly overlapping energy distributions. **d**, Difference spectrogram highlights the mixture regions where speaker one (blue) or two (red) has more acoustic energy. **e**, **f**, Neural-population-based stimulus reconstruction of speaker one (**e**) and speaker two (**f**) alone shows similar spectrotemporal features as the original spectrograms in **a** and **b**. **g**, **h**, The reconstructed spectrograms from the same mixture sound when attending to either speaker one (**g**) or two (**h**) highly resemble the single speaker reconstructions, shown in **e** and **f**, respectively. **i**, Overlay of the spectrogram contours at 50% of maximum energy from the reconstructed spectrograms in **e**, **f**, **g** and **h**.

The average and standard deviation of the correlation between reconstructed and original spectrograms over 24 sentences were 0.60 ± 0.034 (0.60 and 0.62 for the examples in Fig. 1e, f). When attending to each of the two speakers, the reconstructed spectrograms from the same speech mixture showed a marked difference depending upon which speaker was attended (Fig. 1g, h). For each pair, the key temporal and spectral features of the target speaker are enhanced relative to the masker speaker (Fig. 1g, h compared to Fig. 1e, f, respectively). To compare directly, the energy contours from these reconstructed spectrograms are overlaid in Fig. 1i. Important spectrotemporal details of the attended speaker were extracted, while the masker speech was effectively suppressed.

Attentional modulation of the neural representation was quantified, separately for correct and error trials, by measuring the correlation of the reconstructed spectrograms from the mixture in two attended conditions with original acoustic spectrograms of the speakers alone (Fig. 2a–d). During correct trials (Fig. 2a, c), we observed a significant shift of average correlation values towards the target speaker representation. During error trials, in contrast, no significant shift was observed (Fig. 2b, d). Furthermore, the correlations between the reconstructed mixture and the masker speaker were higher than the average intrinsic correlation between randomly chosen original acoustic speech phrases (Fig. 2c, d, dashed lines), revealing a weak presence of the masker speaker in mixture reconstructions, even in correct trials.

The difference in speaking rate of the two speakers, coupled with the stereotyped structure of the carrier phrases, results in specific average temporal modulation profiles for each speaker (average spectrogram for each speaker is shown in Supplementary Fig. 1a, b). To investigate encoding of the distinct spectral profile and characteristic temporal rhythm of the target compared to the masker speaker, we estimated the average difference between reconstructed spectrograms of the two speakers, when presented alone and in the attended mixture (Fig. 2e, f). The comparison between the two average difference reconstructed spectrograms reveals enhanced encoding of both temporal and spectral aspects of the attended speaker (Supplementary Fig. 1c, d). To study the time course of attention-induced modulation of reconstructed mixture spectrograms towards the attended speaker, we

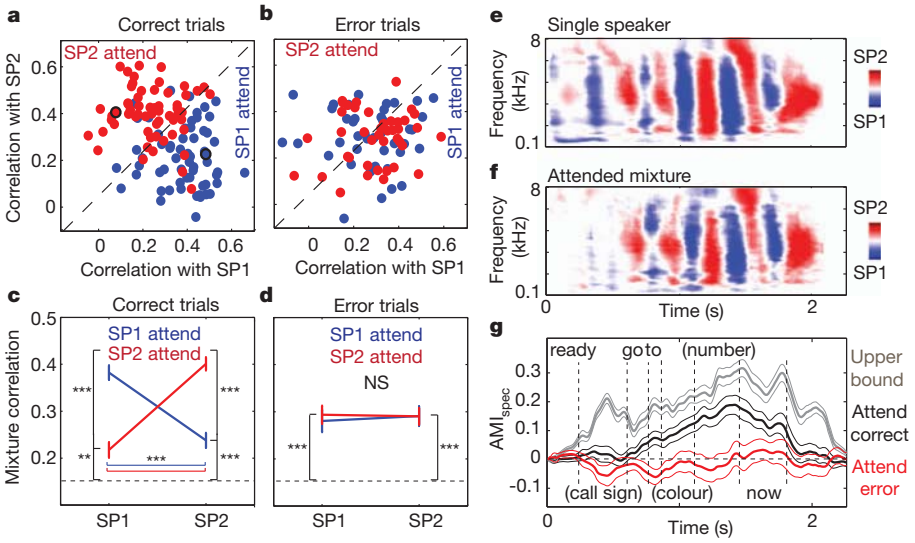


Figure 2 | Quantifying the attentional modulation of neural responses. **a**, **b**, Correlation coefficients of reconstructed mixture spectrograms under attentional control and the corresponding single speaker original spectrograms in correct and error trials (examples in Fig. 1g, h shown with black outline). **c**, **d**, Mean and standard error of correlation values for correct and error trials (28 mixtures). The dashed line corresponds to the average intrinsic correlation between randomly chosen original speech phrases. Brackets indicate pairwise statistical comparisons. NS, not significant. **e**, **f**, Average difference reconstructed spectrograms of speakers one and two from responses to single speaker (**e**) and attended mixture (**f**). **g**, Time course of average and standard error of AMI_{spec} of 28 mixtures for correct (black) and error (red) trials. Grey curve shows the upper bound of AMI_{spec}.

calculated an attentional modulation index (AMI_{spec}), using a sliding window of 250 ms throughout the trial duration:

$$AMI_{spec} = Corr(SP1_{spec}, SP1_{attend}) - Corr(SP1_{spec}, SP2_{attend}) + Corr(SP2_{spec}, SP2_{attend}) - Corr(SP2_{spec}, SP1_{attend}) \quad (1)$$

where $SP1_{spec}$ and $SP2_{spec}$ are the original acoustic spectrograms of speakers one and two, respectively, and $SP1_{attend}$ and $SP2_{attend}$ are the spectrograms reconstructed from neural responses to the mixture with attended targets, speaker one and two, respectively. Positive values of this index reflect shifts towards the target, negative values reflect shifts to the masker representation, and values around zero reflect no shift ($AMI_{spec} = 0.58$ for the example in Fig. 1). An upper bound for the AMI_{spec} was calculated by assuming that attention, at best, restores the single speaker reconstructions of the target speaker (replacing $SP1_{attend}$ and $SP2_{attend}$ in equation (1) with $SP1_{alone}$ and $SP2_{alone}$; Fig. 2g, grey line). The AMI_{spec} from the mixture was first estimated from correct trials (Fig. 2g, black line), and could resolve the time point at which the reconstructed spectrograms were modulated by attention. After the end of the call sign, which cues the speaker that should be attended, a rapid positive shift in the AMI_{spec} was observed, implying the enhanced representation of the target speaker. In error trials, this effect shows a bias towards the masker speaker, which, in contrast, occurred far earlier in the time course. The neural response shift towards the masker, which occurs as early as the call sign, suggests that listeners had prematurely attended to the wrong speaker during those error trials.

Although the reconstruction analyses showed clear attention-based spectrotemporal modulation, we wanted to determine explicitly whether the attended speech in a mixture could be decoded from a model of a single speaker. A regularized linear classifier[21] was trained on neural responses to the single speakers and then used to decode both the spoken words and speaker identity of the attended speech mixture. To keep the chance performance at 50% across all comparisons, classification results were limited only to the choices that were present in each mixture. For correct trials, the colour and number of the attended speech were decoded with high accuracy (77.2% and 80.2%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3a). However, the decoding performance during error trials was significantly below chance (30.0%, 30.1%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3b), indicating a systematic bias towards decoding the words of the masker speaker. In addition, for correct trials, the call sign was classified at chance performance (Fig. 3a). However, for incorrect trials the classifier detected the masker call sign significantly more often than the target call sign (34.1%, $P < 10 \times 10^{-4}$, *t*-test; Fig. 3b), which again shows errors due to an early selection of the masker (incorrect) speaker.

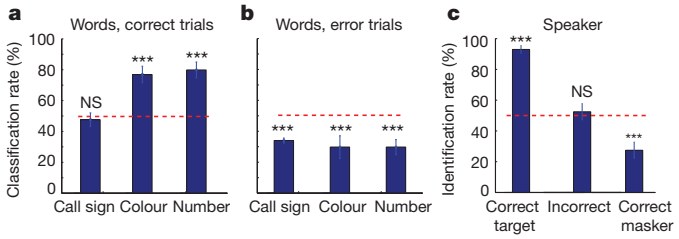


Figure 3 | Decoding spoken words and the identity of the attended speaker. **a**, Classification rate and standard deviation for spoken words (call sign, colour and number) of the attended speaker from the neural responses to the 28 mixtures. Classifiers were trained on single speaker examples only. Colour and number of the attended speech are decoded with high accuracy (77.2% and 80.2%, $P < 10 \times 10^{-4}$, *t*-test) in correct trials, but not the call sign (48.0%, not significant (NS), *t*-test). **b**, In error trials, the classifier showed a systematic bias towards the words of the masker speaker (34.1%, 30.0%, 30.1%, $P < 10 \times 10^{-4}$, *t*-test). **c**, Attended speaker identification rate and standard deviation in correct for target, incorrect (for both target and masker), and correct for masker trials.

For the speaker identification analyses, we divided the behavioural error types into two subsets. The first type occurred when the reported colour–number combination was incorrect for either speaker ('incorrect'; 16.5% of trials). The second type occurred when subjects reported the correct colour–number for the masker instead of the target speaker ('correct for masker'; 8.6% of trials).

In correct trials, the classifier identified the target speaker 93.0% of the time ($P < 10 \times 10^{-4}$, *t*-test; Fig. 3c). During incorrect trials, the classifier performance was at chance. However, during correct for masker trials, the classifier identified the masker rather than the target speaker (27.3%; $P < 10 \times 10^{-4}$, *t*-test; Fig. 3c). These classification results confirm the observed restoration seen in spectrotemporal reconstruction, without necessarily assuming a linear relationship between the neural responses and the stimulus. Furthermore, they extend recent findings using similar methods to decode speech sounds presented in isolation[22] to full words and sentences under complex listening conditions.

We next asked whether the observed robust encoding of attended speech results as an emergent property of the distributed population activity or is driven by a few spatially discrete sites. The cortical regions with reliable evoked responses to speech stimuli were found using a *t*-test between neural responses during speech and silence ($P < 0.01$), and were confined to the posterior superior and middle temporal gyri (Fig. 4a). An example of the attentional response modulation at a single electrode is shown in Fig. 4b–d. The spectrotemporal receptive field (STRF, estimated using the http://www.strflab.berkeley.edu package) of this electrode in passive listening to speech (TIMIT[20]) showed a strong preference for high frequency sounds (Fig. 4b) (STRFs for all electrodes of one subject are provided in Supplementary Fig. 2b). This tuning was also evident in the increased neural response at this electrode (Fig. 4d, dashed lines) to each of the single speakers' high frequency sound components (circled in Fig. 4c, responses are delayed about 120 ms from the stimulus). However, the responses to the same speech mixture sound (Fig. 4d, solid lines) were significantly modulated by attention. The responses to high frequency components were enhanced for the attended speaker, but suppressed for similar sounds in the masker speaker (Fig. 4d, solid lines compared to dashed lines). This highly modulated yet fixed feature selectivity probably contributes to the constancy of the single speaker representation observed in our previous analyses. To quantify this effect for each individual electrode, we measured the correlation between the neural responses to the attended mixture and to those of the speakers in isolation (AMI_{elec}, equation (2) in Methods). We found a varying degree of bias towards the attended speaker distributed across the population (Supplementary Fig. 3d; $AMI_{elec} = 0.28$ for the example

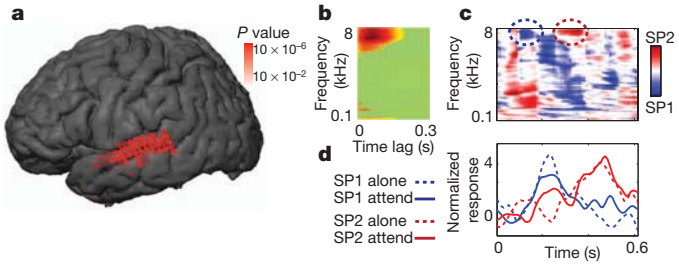


Figure 4 | Attentional modulation of individual electrode sites. **a**, Electrodes picking up a significant difference between responses to silence and speech sounds ($P < 0.01$, *t*-test). **b**, STRF of this representative electrode site shows a preference for high frequency sounds. **c**, Mixture difference spectrogram for a selected duration containing a high frequency component for each speaker (circled). **d**, The electrode shows an increased response to high frequency sounds of single speakers (dashed lines, peak neural response is delayed by about 120 ms). However, the neural response to the same mixture sound in two attention conditions (solid lines) showed an enhanced response to high frequency sounds only for the target, but with responses for similar sounds in the masker speaker suppressed.

in Fig. 4), which gradually builds up after the end of the call sign (Supplementary Fig. 3e). We did not observe any particular anatomical pattern for the attentional modulation across sites (Supplementary Fig. 3f). Rather, it appeared to be distributed over responsive sites, consistent with previous findings of higher-order sound processing[23].

In summary, we demonstrate that the human auditory system restores the representation of the attended speaker while suppressing irrelevant competing speech. Speech restoration occurs at a level where neural responses still show precise phase-locking to spectrotemporal features of speech. Population responses revealed the emergent representation of speech extracted from a mixture, including the moment-by-moment allocation of attentional focus.

These results have implications for models of auditory scene analysis. In agreement with recent studies, the cortical representation of speech in the posterior temporal lobe does not merely reflect the acoustical properties of the stimulus, but instead relates strongly to the perceived aspects of speech[10]. Although the exact mechanisms are not fully known, multiple processes in addition to attention are likely to enable this high-order auditory processing, including grouping of predictable regularities in speech acoustics[24], feature binding[3,25] and phonemic restoration[26]. Conversely, behavioural errors seem to result from degradation of the neural representation, a direct result of inherent sensory interference such as energetic masking[16] (Supplementary Fig. 3g, h) and/or the allocation of attention[27].

In speech, the end result represented in the posterior temporal lobe appears to be unaffected by perceptually irrelevant sounds, which is ideal for subsequent linguistic and cognitive processing. Following one speaker in the presence of another can be trivial for a normal human listener, but remains a major challenge for state-of-the-art automatic speech recognition algorithms[28]. Understanding how the brain solves this problem may inspire more efficient and generalizable solutions than current engineering approaches[29]. It will also shed light on how these processes become impaired during ageing and in disorders of speech perception in real-world hearing conditions[7].

METHODS SUMMARY

Three human subjects with normal hearing underwent the placement of a subdural electrode array as part of their clinical treatment for epilepsy. We used speech samples from a publicly available database called Coordinate Response Measure (CRM[15]). One male and one female speaker were selected with two call signs (ringo and tiger), three colours (red, blue or green) and three numbers (two, five or seven). We generated 12 unique combinations of call sign, colour and number per speaker (total of 24 single speaker phrases) and 28 mixture speech samples by selecting from combinations of the 24 single speaker sentences (0 dB target-to-masker ratio). Speech sounds were presented monaurally from a loud speaker. We used stimulus reconstruction[17–19] to map the population electrocorticographic response to the spectrogram of the speech stimulus. Reconstruction filters were estimated from neural responses to a separate speech corpus (TIMIT[20]). Test speakers were not used in the estimation of filters. For word and speaker decoding analysis, a regularized linear classifier[21] was trained on neural responses of the single speakers and then used to decode the spoken words and speaker identity of the attended speech mixture.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 30 August 2011; accepted 5 March 2012.
Published online 18 April 2012.

1. Cherry, E. C. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **25,** 975–979 (1953).
2. Shinn-Cunningham, B. G. Object-based auditory and visual attention. *Trends Cogn. Sci.* **12,** 182–186 (2008).
3. Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, 1994).
4. Kerlin, J., Shahin, A. & Miller, L. Attentional gain control of ongoing cortical speech representations in a ''cocktail party''. *J. Neurosci.* **30,** 620–628 (2010).
5. Besle, J. *et al.* Tuning of the human neocortex to the temporal dynamics of attended events. *J. Neurosci.* **31,** 3176–3185 (2011).
6. Bee, M. & Micheyl, C. The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? *J. Comparative Psychol.* **122,** 235–252 (2008).
7. Shinn-Cunningham, B. G. & Best, V. Selective attention in normal and impaired hearing. *Trends Amplif.* **12,** 283–299 (2008).
8. Scott, S. K., Rosen, S., Beaman, C. P., Davis, J. P. & Wise, R. J. S. The neural processing of masked speech: evidence for different mechanisms in the left and right temporal lobes. *J. Acoust. Soc. Am.* **125,** 1737–1743 (2009).
9. Elhilali, M., Xiang, J., Shamma, S. A. & Simon, J. Z. Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene. *PLoS Biol.* **7,** e1000129 (2009).
10. Chang, E. F. *et al.* Categorical speech representation in human superior temporal gyrus. *Nature Neurosci.* **13,** 1428–1432 (2010).
11. Crone, N. E., Boatman, D., Gordon, B. & Hao, L. Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* **112,** 565–582 (2001).
12. Steinschneider, M., Fishman, Y. I. & Arezzo, J. C. Spectrotemporal analysis of evoked and induced electroencephalographic responses in primary auditory cortex (A1) of the awake monkey. *Cereb. Cortex* **18,** 610–625 (2008).
13. Scott, S. K. & Johnsrude, I. S. The neuroanatomical and functional organization of speech perception. *Trends Neurosci.* **26,** 100–107 (2003).
14. Hackett, T. A. Information flow in the auditory cortical network. *Hear. Res.* **271,** 133–146 (2011).
15. Bolia, R. S., Nelson, W. T., Ericson, M. A. & Simpson, B. D. A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* **107,** 1065–1066 (2000).
16. Brungart, D. S. Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109,** 1101–1109 (2001).
17. Mesgarani, N., David, S. V., Fritz, J. B. & Shamma, S. A. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J. Neurophysiol.* **102,** 3329–3339 (2009).
18. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252,** 1854–1857 (1991).
19. Pasley, B. N. *et al.* Reconstructing speech from human auditory cortex. *PLoS Biol.* **10,** e1001251 (2012).
20. Garofolo, J. S. *et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus* (Linguistic Data Consortium, 1993).
21. Rifkin, R., Yeo, G. & Poggio, T. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences* **190,** 131–154 (2003).
22. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. ''Who'' is saying ''what''? Brain-based decoding of human voice and speech. *Science* **322,** 970–973 (2008).
23. Staeren, N., Renvall, H., De Martino, F., Goebel, R. & Formisano, E. Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* **19,** 498–502 (2009).
24. Shamma, S. A., Elhilali, M. & Micheyl, C. Temporal coherence and attention in auditory scene analysis. *Trends Neurosci.* **34,** 114–123 (2010).
25. Darwin, C. J. Auditory grouping. *Trends Cogn. Sci.* **1,** 327–333 (1997).
26. Warren, R. M. Perceptual restoration of missing speech sounds. *Science* **167,** 392–393 (1970).
27. Kidd, G. Jr, Arbogast, T. L., Mason, C. R. & Gallun, F. J. The advantage of knowing where to listen. *J. Acoust. Soc. Am.* **118,** 3804–3815 (2005).
28. Shen, W., Olive, J. & Jones, D. Two protocols comparing human and machine phonetic discrimination performance in conversational speech. *INTERSPEECH* 1630–1633 (2008).
29. Cooke, M., Hershey, J. R. & Rennie, S. J. Monaural speech separation and recognition challenge. *Comput. Speech Lang.* **24,** 1–15 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank A. Ren for technical help, and C. Micheyl, S. Shamma and C. Schreiner for critical discussion and reading of the manuscript. E.F.C. was funded by National Institutes of Health grants R00-NS065120, DP2-OD00862, R01-DC012379, and the Ester A. and Joseph Klingenstein Foundation.

Author Contributions N.M. and E.F.C. designed the experiment, collected the data, evaluated results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.F.C. (changed@neurosurg.ucsf.edu).

METHODS

The experimental protocol was approved by the Committee for Human Research at the University of California, San Francisco.

Subjects. Three human subjects underwent the placement of a high-density subdural electrode array (4 mm pitch) over the language-dominant hemisphere as part of routine clinical treatment for epilepsy. Subjects gave their written informed consent before surgery. All subjects had self-reported normal hearing and underwent neuropsychological language testing (including the Boston naming and verbal fluency tests) and were found to be normal. The intracarotid sodium amobarbital (Wada) test was used for language dominance assessment. The electrodes in the study were located over the posterior dorsolateral temporal lobe. The location and corresponding spectrotemporal receptive fields of all the included electrodes for a subject are shown in Supplementary Fig. 2.

Data acquisition and pre-processing. The electrocorticography signal was recorded with a multichannel amplifier optically connected to a digital signal processor (TuckerDavis Technologies). Each channel time series was visually and quantitatively inspected for artefacts or excessive noise. The data were then segmented with a 100 ms pre-stimulus baseline and a 400 ms post-stimulus interval. The common mode signal was estimated using principal component analysis with channels as repetitions and was removed from each channel time series using vector projection.

Task design and behavioural testing. We used speech samples from a publicly available database called Coordinate Response Measure (CRM[15]) containing sentences in the form "ready (call sign) go to (colour) (number) now". One male and one female speaker (speakers one and five in CRM corpus) were selected with two call signs (ringo and tiger), three colours (blue (B), red (R) or green (G)) and three numbers (two, five or seven). For each of the two call signs, we generated six colour–number combinations (B2, B5, R2, R7, G5, G7), resulting in 12 different phrases. We chose the same phrases for each of the two speakers, resulting in 24 single speaker sentences. We then produced 28 unique mixture speech samples by selecting from combinations of the 24 single speaker sentences at 0 dB target-to-masker ratio. Each mixture sample was chosen such that there was no overlap between call signs, colours or the numbers of the two phrases. In addition, each speaker had the same number of call signs (ringo or tiger) in each trial block. The sounds were presented monaurally from a loudspeaker connected to a laptop, which was also used to collect subjects' responses through a customized graphical user interface. Each trial block consisted of 28 trials and the target call sign was fixed for each block. The target call sign was displayed visually before and during the trial block. Subjects first listened to each of the speakers alone and were able to report the colour and number with 100% accuracy. Subjects then listened to a monaural, simultaneous mixture of the two speakers' phrases with different call signs, colours and numbers. The subjects were instructed to respond by indicating the colour and number spoken by the talker who uttered the target call sign. The target speaker changed from trial to trial pseudo-randomly, requiring the subjects to initially monitor both speakers until they detect the target call sign. After each trial block, the target call sign was changed, switching the role of target and masker speakers in each mixture sound.

Electrode selection. The cortical sites on the superior and middle temporal gyri with reliable evoked responses to speech stimuli were selected for all the subsequent analysis. Our inclusion criteria consisted of a *t*-test between responses to randomly selected time frames during passive speech presentation (TIMIT) and in silence ($P < 0.01$, resulting in 83, 92 and 102 electrodes for subjects one to three. One example subject is shown in Supplementary Fig. 2a). Solely for visualization, we also estimated the STRFs of these selected sites from passive listening to TIMIT using normalized reverse correlation algorithm (STRFLab software package, http://www.strflab.berkeley.edu; Supplementary Fig. 2b). Correlation histogram of STRF predictions for all 275 electrode sites is shown in Supplementary Fig. 1c.

Stimulus reconstruction. We used stimulus reconstruction to map the population neural responses to the spectrogram of the speech stimulus[17–19]. Reconstruction filters were estimated from neural responses to a separate speech corpus (TIMIT[20]) containing a total of 499 unique short sentences from 402 different speakers. Filters were obtained using normalized reverse correlation to minimize the mean squared error of the reconstructed spectrograms[17] with filter time lags from −420 to 0 ms (causal filters). The filters were then fixed in all subsequent conditions and were applied to the neural responses to CRM samples. Neither of the speakers or phrases in the CRM data set was used in estimation of the filters. The output of the reconstruction algorithm was further processed with a band-pass filter applied to each frequency channel of reconstructed spectrograms to remove the baseline. All the processing steps for stimulus reconstruction were identical in all conditions (single and mixture speakers).

AMI. To quantify the change in similarity between the representation of single and attended speaker in mixture speech, we defined the AMI_{spec} in equation (1). The stereotypical format of the CRM phrases results in an intrinsic correlation between the neural responses to different sentences, particularly at the beginning ("ready") and middle of the carrier phrase ("go to"), which results in reduced possible AMI_{spec} values for these segments. To estimate an upper bound for unbiased comparison, AMI_{spec} was calculated where the representation of an attended speaker in a mixture is ideally assumed to be identical to the representation of that speaker when presented alone; therefore, replacing SP_{attend} in equation (1) with the reconstructed spectrogram of single speaker SP_{alone}. The upper bound peaks at the call sign, colour and number where different phrases are most dissimilar. The overall increase in the upper bound is due to the progressive asynchrony between the two speakers.

The same statistics can be used to estimate the AMI of an individual electrode site by calculating the correlation values between the neural response of that site to attended mixture and single speaker presentations:

$$\begin{aligned} AMI_{elec} = {} & Corr(\text{R-SP1}_{alone}, \text{R-SP1}_{attend}) - Corr(\text{R-SP1}_{alone}, \text{R-SP2}_{attend}) + \\ & Corr(\text{R-SP2}_{alone}, \text{R-SP2}_{attend}) - Corr(\text{R-SP2}_{alone}, \text{R-SP1}_{attend}) \end{aligned} \quad (2)$$

where R-SP1_{alone} and R-SP2_{alone} are the responses of an electrode to speakers one and two alone, respectively, and R-SP1_{attend} and R-SP2_{attend} are the responses of the same electrode to the mixture of the two when the attended target is speaker one and two, respectively.

Classification of spoken words and speaker identity. A linear-frame-based regularized-least-square classifier[21] was used to investigate the discriminability of the spoken words and speaker identity from electrocorticographic responses. Two binary classifiers were trained to classify the call sign and speaker identity, and two separate three-way classifiers were used for colour and for number classification. Classifiers were trained only on the neural responses of single speakers (24 sentences) and tested on the mixtures. The classifiers produced a linear weighted sum of the neural responses at each time instance and the classifier that produced the maximum average output over the duration of words was chosen as classification result. The classifier decision was limited to only the colours and numbers that occurred in each mixture, therefore resulting in same 50% chance performance in all cases.

LETTER

doi:10.1038/nature10945

De novo mutations revealed by whole-exome sequencing are strongly associated with autism

Stephan J. Sanders[1], Michael T. Murtha[1], Abha R. Gupta[2]*, John D. Murdoch[1]*, Melanie J. Raubeson[1]*, A. Jeremy Willsey[1]*, A. Gulhan Ercan-Sencicek[1]*, Nicholas M. DiLullo[1]*, Neelroop N. Parikshak[3], Jason L. Stein[3], Michael F. Walker[1], Gordon T. Ober[1], Nicole A. Teran[1], Youeun Song[1], Paul El-Fishawy[1], Ryan C. Murtha[1], Murim Choi[4], John D. Overton[4], Robert D. Bjornson[5], Nicholas J. Carriero[5], Kyle A. Meyer[6], Kaya Bilguvar[7], Shrikant M. Mane[8], Nenad Šestan[6], Richard P. Lifton[4], Murat Günel[7], Kathryn Roeder[9], Daniel H. Geschwind[3], Bernie Devlin[10] & Matthew W. State[1]

Multiple studies have confirmed the contribution of rare *de novo* copy number variations to the risk for autism spectrum disorders[1–3]. But whereas *de novo* single nucleotide variants have been identified in affected individuals[4], their contribution to risk has yet to be clarified. Specifically, the frequency and distribution of these mutations have not been well characterized in matched unaffected controls, and such data are vital to the interpretation of *de novo* coding mutations observed in probands. Here we show, using whole-exome sequencing of 928 individuals, including 200 phenotypically discordant sibling pairs, that highly disruptive (nonsense and splice-site) *de novo* mutations in brain-expressed genes are associated with autism spectrum disorders and carry large effects. On the basis of mutation rates in unaffected individuals, we demonstrate that multiple independent *de novo* single nucleotide variants in the same gene among unrelated probands reliably identifies risk alleles, providing a clear path forward for gene discovery. Among a total of 279 identified *de novo* coding mutations, there is a single instance in probands, and none in siblings, in which two independent nonsense variants disrupt the same gene, *SCN2A* (sodium channel, voltage-gated, type II, α subunit), a result that is highly unlikely by chance.

We completed whole-exome sequencing in 238 families from the Simons Simplex Collection (SSC), a comprehensively phenotyped autism spectrum disorders (ASD) cohort consisting of pedigrees with two unaffected parents, an affected proband, and, in 200 families, an unaffected sibling[5]. Exome sequences were captured with NimbleGen oligonucleotide libraries, subjected to DNA sequencing on the Illumina platform, and genotype calls were made at targeted bases (Supplementary Information)[6,7]. On average, 95% of the targeted bases in each individual were assessed by ≥8 independent sequence reads; only those bases showing ≥20 independent reads in all family members were considered for *de novo* mutation detection. This allowed for analysis of *de novo* events in 83% of all targeted bases and 73% of all exons and splice sites in the RefSeq hg18 database (http://www.ncbi.nlm.nih.gov/RefSeq/; Supplementary Table 1; Supplementary Data 1). Given uncertainties regarding the sensitivity of detection of insertion-deletions, case-control comparisons reported here consider only single base substitutions (Supplementary Information). Validation was attempted for all predicted *de novo* single nucleotide variants (SNVs) via Sanger sequencing of all family members, with sequence readers blinded to affected status; 96% were successfully validated. We determined there was no evidence of systematic bias in variant detection between affected and unaffected siblings through comparisons of silent *de novo*, non-coding *de novo*, and novel transmitted variants (Fig. 1a; Supplementary Figs 1–5; Supplementary Information).

Among 200 quartets (Table 1), 125 non-synonymous *de novo* SNVs were present in probands and 87 in siblings: 15 of these were nonsense (10 in probands; 5 in siblings) and 5 altered a canonical splice site (5 in probands; 0 in siblings). There were 2 instances in which *de novo* SNVs were present in the same gene in two unrelated probands; one of these involved two independent nonsense variants (Table 2). Overall, the total number of non-synonymous *de novo* SNVs was significantly greater in probands compared to their unaffected siblings ($P = 0.01$, two-tailed binomial exact test; Fig. 1a; Table 1) as was the odds ratio (OR) of non-synonymous to silent mutations in probands versus siblings (OR = 1.93; 95% confidence interval (CI), 1.11–3.36; $P = 0.02$, asymptotic test; Table 1). Restricting the analysis to nonsense and splice site mutations in brain-expressed genes resulted in substantially increased estimates of effect size and demonstrated a significant difference in cases versus controls based either on an analysis of mutation burden ($N = 13$ versus 3; $P = 0.02$, two-tailed binomial exact test; Fig. 1a; Table 1) or an evaluation of the odds ratio of nonsense and splice site to silent SNVs (OR = 5.65; 95% CI, 1.44–22.2; $P = 0.01$, asymptotic test; Fig. 1b; Table 1).

To determine whether factors other than diagnosis of ASD could explain our findings, we examined a variety of potential covariates, including parental age, IQ and sex. We found that the rate of *de novo* SNVs indeed increases with paternal age ($P = 0.008$, two-tailed Poisson regression) and that paternal and maternal ages are highly correlated ($P < 0.0001$, two-tailed linear regression). However, although the mean paternal age of probands in our sample was 1.1 years higher than their unaffected siblings, re-analysis accounting for age did not substantively alter any of the significant results reported here (Supplementary Information). Similarly, no significant relationship was observed between the rate of *de novo* SNVs and proband IQ ($P \geq 0.19$, two-tailed linear regression, Supplementary Information) or proband sex ($P \geq 0.12$, two-tailed Poisson regression; Supplementary Fig. 6; Supplementary Information).

Overall, these data demonstrate that non-synonymous *de novo* SNVs, and particularly highly disruptive nonsense and splice-site *de novo* mutations, are associated with ASD. On the basis of the conservative assumption that *de novo* single-base coding mutations observed in siblings confer no autism liability, we estimate that at least 14% of

[1]Program on Neurogenetics, Child Study Center, Department of Psychiatry, Department of Genetics, Yale University School of Medicine, 230 South Frontage Road, New Haven, Connecticut 06520, USA. [2]Child Study Center, Department of Pediatrics, Yale University School of Medicine, 230 South Frontage Road, New Haven, Connecticut 06520, USA. [3]Neurogenetics Program, UCLA, 695 Charles E. Young Dr. South, Los Angeles, California 90095, USA. [4]Department of Genetics, Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, Connecticut 06510, USA. [5]Department of Computer Science, Yale Center for Genome Analysis, Yale University, 51 Prospect Street, New Haven, Connecticut 06511, USA. [6]Department of Neurobiology, Kavli Institute for Neuroscience, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA. [7]Department of Neurosurgery, Center for Human Genetics and Genomics, Program on Neurogenetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06520, USA. [8]Yale Center for Genome Analysis, 300 Heffernan Drive, West Haven, Connecticut 06516, USA. [9]Department of Statistics, Carnegie Mellon University, 130 DeSoto Street, Pittsburgh, Pennsylvania 15213, USA. [10]Department of Psychiatry and Human Genetics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. *These authors contributed equally to this work.

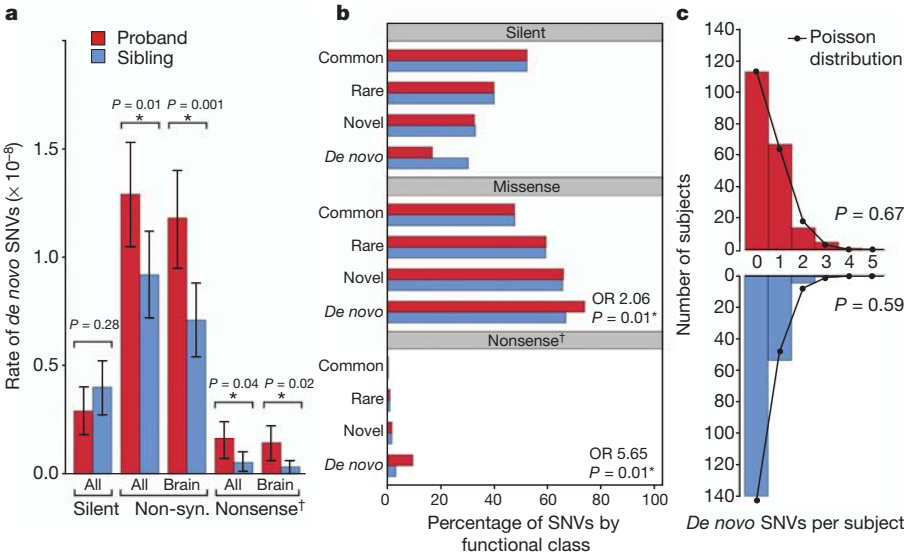


Figure 1 | Enrichment of non-synonymous *de novo* variants in probands relative to sibling controls. **a**, The rate of *de novo* variants is shown for 200 probands (red) and matched unaffected siblings (blue). 'All' refers to all RefSeq genes in hg18, 'Brain' refers to the subset of genes that are brain-expressed[24] and 'Non-syn' to non-synonymous SNVs (including missense, nonsense and splice site SNVs). Error bars represent the 95% confidence intervals and *P* values are calculated with a two-tailed binomial exact test. **b**, The proportion of transmitted variants in brain-expressed genes is equal between 200 probands (red) and matched unaffected siblings (blue) for all mutation types and allele frequencies, including common (≥1%), rare (<1%) and novel (single allele in one of the 400 parents); in contrast, both non-synonymous and nonsense *de novo* variants show significant enrichment in probands compared to unaffected siblings (73.7% versus 66.7%, $P = 0.01$, asymptotic test and 9.5% versus 3.1%, $P = 0.01$ respectively). **c**, The frequency distribution of brain-expressed non-synonymous *de novo* SNVs is shown per sample for probands (red) and siblings (blue). Neither distribution differs from the Poisson distribution (black line), suggesting that multiple *de novo* SNVs within a single individual do not confirm ASD risk. Nonsense† represents the combination of nonsense and splice site SNVs.

affected individuals in the SSC carry *de novo* SNV risk events (Supplementary Information). Moreover, among probands and considering brain-expressed genes, an estimated 41% of non-synonymous *de novo* SNVs (95% CI, 21–58%) and 77% of nonsense and splice site *de novo* SNVs (95% CI, 33–100%) point to *bona fide* ASD-risk loci (Supplementary Information).

We next set out to evaluate which of the particular *de novo* SNVs identified in our study confer this risk. On the basis of our prior work[3],

Table 1 | Distribution of SNVs between probands and siblings

Category	Total number of SNVs*		SNVs per subject		Per base SNV rate (x10^{-8})		*P*†	Odds ratio (95% CI)‡
	Pro $N = 200$	Sib $N = 200$	Pro $N = 200$	Sib $N = 200$	Pro $N = 200$	Sib $N = 200$		
De novo								
				All genes				
All	154	125 §	0.77	0.63	1.58	1.31	0.09	NA
Silent	29	39	0.15	0.20	0.29	0.40	0.28	NA
All non-synonymous	125	87	0.63	0.44	1.29	0.92	**0.01**	1.93 (1.11–3.36)
Missense	110	82	0.55	0.41	1.13	0.86	0.05	1.80 (1.03–3.16)
Nonsense/splice site	15	5	0.08	0.03	0.16	0.05	**0.04**	4.03 (1.32-12.4)
				Brain-expressed genes				
All	137	96	0.69	0.48	1.41	1.01	**0.01**	NA
Silent	23	30	0.12	0.15	0.24	0.31	0.41	NA
All non-synonymous	114	67	0.57	0.34	1.18	0.71	**0.001**	2.22 (1.19–4.13)
Missense	101	64	0.51	0.32	1.04	0.68	**0.005**	2.06 (1.10–3.85)
Nonsense/splice site	13	3	0.07	0.02	0.14	0.03	**0.02**	5.65 (1.44–22.2)
Novel transmitted								
				All genes				
All	26,565	26,542	133	133	277	277	0.92	NA
Silent	8,567	8,642	43	43	90	91	0.57	NA
All non-synonymous	17,998	17,900	90	90	188	187	0.61	1.01 (0.98–1.05)
Missense	17,348	17,250	87	86	181	180	0.60	1.01 (0.98–1.05)
Nonsense/splice site	650	650	3.3	3.3	7	7	1.00	1.01 (0.90–1.13)
				Brain-expressed genes				
All	20,942	20,982	105	105	219	220	0.85	NA
Silent	6,884	6,981	34	35	72	74	0.42	NA
All non-synonymous	14,058	14,001	70	70	147	146	0.74	1.02 (0.98–1.06)
Missense	13,588	13,525	68	68	142	141	0.71	1.02 (0.98–1.06)
Nonsense/splice site	470	476	2.3	2.4	5	5	0.87	1.00 (0.88–1.14)

* An additional 15 *de novo* variants were seen in the probands of 25 trio families; all were missense and 14 were brain-expressed.
† The *P* values compare the number of variants between probands and siblings using a two-tailed binomial exact test (Supplementary Information); *P* values below 0.05 are highlighted in bold.
‡ The odds ratio calculates the proportion of variants in a specific category to silent variants and then compares these ratios in probands versus siblings. NA, not applicable.
§ The sum of silent and non-synonymous variants is 126, however one nonsense and two silent *de novo* variants were indentified in *KANK1* in a single sibling, suggesting a single gene conversion event. This event contributed a maximum count of one to any analysis.

Table 2 | Loss of function mutations in probands

Gene symbol	Gene name	Mutation type
ADAM33	ADAM metallopeptidase domain 33	Nonsense
CSDE1	cold shock domain containing E1, RNA-binding	Nonsense
EPHB2	EPH receptor B2	Nonsense
FAM8A1	family with sequence similarity 8, member A1	Nonsense
FREM3	FRAS1 related extracellular matrix 3	Nonsense
MPHOSPH8	M-phase phosphoprotein 8	Nonsense
PPM1D	protein phosphatase, Mg^{2+}/Mn^{2+} dependent 1D	Nonsense
RAB2A	RAB2A, member RAS oncogene family	Nonsense
SCN2A	sodium channel, voltage-gated, type II, α subunit	Nonsense
SCN2A	sodium channel, voltage-gated, type II, α subunit	Nonsense
BTN1A1	butyrophilin, subfamily 1, member A1	Splice site
FCRL6	Fc receptor-like 6	Splice site
KATNAL2	katanin p60 subunit A-like 2	Splice site
NAPRT1	nicotinate phosphoribosyltransferase domain containing 1	Splice site
RNF38	ring finger protein 38	Splice site
SCP2	sterol carrier protein 2	Frameshift*
SHANK2	SH3 and multiple ankyrin repeat domains 2	Frameshift*

* Frameshift *de novo* variants are not included in any of the reported case-control comparisons (Supplementary Information).

we hypothesized that estimating the probability of observing multiple independent *de novo* SNVs in the same gene in unrelated individuals would provide a more powerful statistical approach to identifying ASD-risk genes than the alternative of comparing mutation counts in affected versus unaffected individuals. Consequently, we conducted simulation experiments focusing on *de novo* SNVs in brain-expressed genes, using the empirical data for per-base mutation rates and taking into account the actual distribution of gene sizes and GC content across the genome (Supplementary Information). We calculated probabilities (P) and the false discovery rate (Q) based on a wide range of assumptions regarding the number of genes conferring ASD risk (Supplementary Fig. 7; Fig. 2). On the basis of 150,000 iterations, we determined that under all models, two or more nonsense and/or splice site *de novo* mutations were highly unlikely to occur by chance ($P = 0.008$; $Q = 0.005$; Supplementary Information; Fig. 2a). Importantly, these thresholds were robust both to sample size, and to variation in our estimates of locus heterogeneity. Similarly, in our sample, two or more nonsense or splice site *de novo* mutations remained statistically significant when the simulation was performed using the lower bound of the 95% confidence interval for the estimate of *de novo* mutation rates in probands (Supplementary Fig. 7).

Only a single gene in our cohort, *SCN2A*, met these thresholds ($P = 0.008$; Fig. 2a), with two probands each carrying a nonsense *de novo* SNV (Table 2). This finding is consistent with a wealth of data showing overlap of genetic risks for ASD and seizure[8]. Gain of function mutations in *SCN2A* are associated with a range of epilepsy phenotypes; a nonsense *de novo* mutation has been described in a patient with infantile epileptic encephalopathy and intellectual decline[9], *de novo* missense mutations with variable electrophysiological effects have been found in cases of intractable epilepsy[10], and transmitted rare missense mutations have been described in families with idiopathic ASD[11]. Of note, the individuals in the SSC carrying the nonsense *de novo* SNVs have no history of seizure.

We then considered whether alternative approaches described in the recent literature[4,12], including identifying multiple *de novo* events in a single individual or predicting the functional consequences of missense mutations, might help identify additional ASD-risk genes. However, we found no differences in the distribution or frequency of multiple *de novo* events within individuals in the case versus the control groups (Fig. 1c). In addition, when we examined patients carrying large *de novo* ASD-risk CNVs, we found a trend towards fewer non-synonymous *de novo* SNVs (Supplementary Fig. 11; Supplementary Information). Consequently, neither finding supported a 'two *de novo* hit' hypothesis. Similarly, we found no evidence that widely used measures of conservation or predictors of protein disruption, such as PolyPhen2[13], SIFT[14], GERP[15], PhyloP[16] or Grantham Score[17],

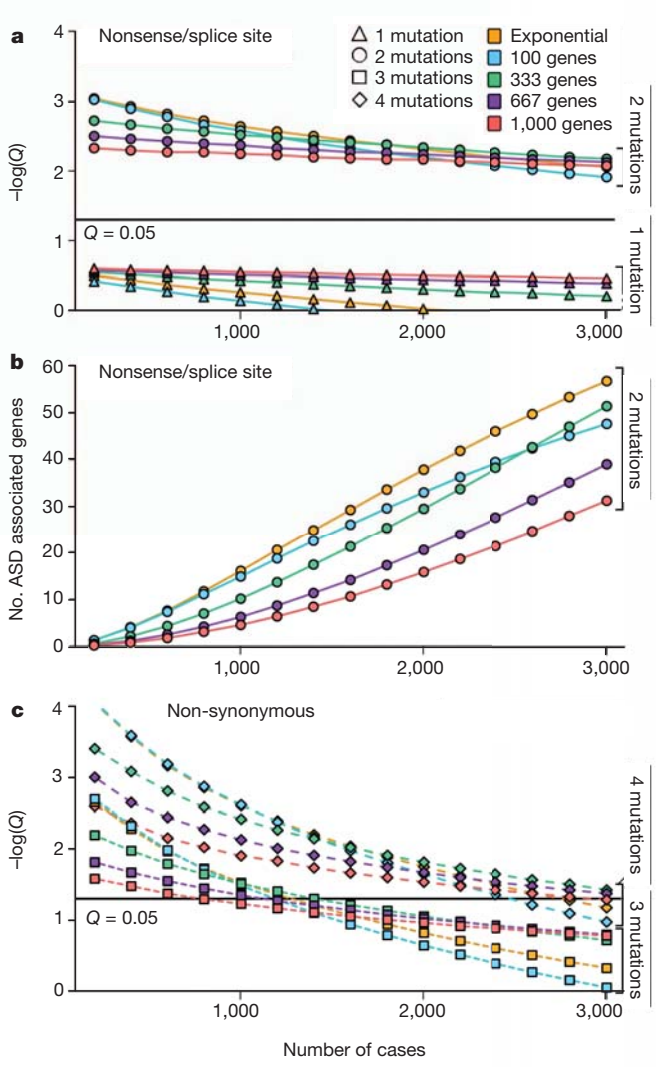


Figure 2 | Identification of multiple *de novo* mutations in the same gene reliably distinguishes risk-associated mutations. **a**, Results of a simulation experiment modelling the likelihood of observing two independent nonsense/splice site *de novo* mutations in the same brain-expressed gene among unrelated probands. We modelled the observed rate of *de novo* brain-expressed mutations in probands and siblings, gene size, GC content and varying degrees of locus heterogeneity, including 100, 333, 667 or 1,000 ASD-contributing genes, as well as using the top 1% of genes derived from a model of exponential distribution of risk (indicated by colour). A total of 150,000 iterations were run. The rate of occurrences of two or more *de novo* variants in non-ASD genes was used to estimate the *P*-value (Supplementary Fig. 7) while the ratio of occurrences of two or more *de novo* variants in non-ASD genes to similar occurrences in ASD genes was used to estimate the false discovery rate (Q). The identification of two independent nonsense/splice site *de novo* variants in a brain-expressed gene in this sample provides significant evidence for ASD association ($P = 0.008$; $Q = 0.005$) for all models. This observation remained statistically significant when the simulation was repeated using the lower bound of the 95% confidence interval for the estimate of the *de novo* mutation rate in probands (Supplementary Fig. 7). **b**, The simulation described in **a** was used to predict the number of genes that will be found to carry two or more nonsense/splice site *de novo* mutations for a sample of a given size (specified on the *x* axis). **c**, The simulation was repeated for non-synonymous *de novo* mutations. The identification of three or more independent non-synonymous *de novo* mutations in a brain-expressed gene provides significant evidence for ASD association ($P < 0.05$; $Q < 0.05$) in the sample reported here, however these thresholds are sensitive both to sample size and heterogeneity models.

either alone or in combination differentiated *de novo* non-synonymous SNVs in probands compared to siblings (Supplementary Fig. 9; Supplementary Information). Additionally, among probands, the *de*

novo SNVs in our study were not significantly over-represented in previously established lists of synaptic genes[18–20], genes on chromosome X, autism-implicated genes[2], intellectual disability genes[2], genes within ASD-risk associated CNVs[3] or *de novo* non-synonymous SNVs identified in schizophrenia probands[12,21]. Finally we conducted pathway and protein–protein interaction analyses[22] for all non-synonymous *de novo* SNVs, all brain-expressed non-synonymous *de novo* SNVs and all nonsense and splice site *de novo* SNVs (Supplementary Fig. 9, 10; Supplementary Information) and did not find a significant enrichment among cases versus controls that survived correction for multiple comparisons, though these studies were of limited power.

These analyses demonstrate that neither the type nor the number of *de novo* mutations observed solely in a single individual provides significant evidence for association with ASD. Moreover, we determined that in the SSC cohort at least three, and most often four or more, brain-expressed non-synonymous *de novo* SNVs in the same gene would be necessary to show a significant association (Fig. 2c; Supplementary Figs 7, 8). Unlike the case of disruptive nonsense and splice site mutations, these simulations were highly sensitive to both sample size and heterogeneity models (Fig. 2c; Supplementary Figs 7, 8; Supplementary Information).

Finally, at the completion of our study, we had the opportunity to combine all *de novo* events in our sample with those identified in an independent whole-exome analysis of non-overlapping Simons Simplex families that focused predominantly on trios[23]. From a total of 414 probands, two additional genes were found to carry two highly disruptive mutations each, *KATNAL2* (katanin p60 subunit A-like 2) (our results and ref. 23) and *CHD8* (chromodomain helicase DNA binding protein 8) (ref. 23), thereby showing association with the ASD phenotype.

Overall, our results substantially clarify the genomic architecture of ASD, demonstrate significant association of three genes—*SCN2A*, *KATNAL2* and *CHD8*—and predict that approximately 25–50 additional ASD-risk genes will be identified as sequencing of the 2,648 SSC families is completed (Fig. 2b). Rare non-synonymous *de novo* SNVs are associated with risk, with odds ratios for nonsense and splice-site mutations in the range previously described for large multigenic *de novo* CNVs[3]. It is important to note that these estimates reflect a mix of risk and neutral mutations in probands. We anticipate that the true effect size for specific SNVs and mutation classes will be further clarified as more data accumulate. From the distribution of large multi-genic *de novo* CNVs in probands versus siblings, we previously estimated the number of ASD-risk loci at 234 (ref. 3). Using the same approach, the current data result in a point estimate of 1,034 genes, however the confidence intervals are large and the distribution of this risk among these loci is unknown (Supplementary Information). What is clear is that our results strongly support a high degree of locus heterogeneity in the SSC cohort, involving hundreds of genes or more. Finally, via examination of mutation rates in well-matched controls, we have determined that the observation of highly disruptive *de novo* SNVs clustering within genes can robustly identify risk-conferring alleles.

The focus on recurrent rare *de novo* mutation described here provided sufficient statistical power to identify associated genes in a relatively small cohort—despite both a high degree of locus heterogeneity and the contribution of intermediate genetic risks. This approach promises to be valuable for future high-throughput sequencing efforts in ASD and other common neuropsychiatric disorders.

METHODS SUMMARY

Sample selection. In total 238 families (928 individuals) were selected from the SSC[5]. Thirteen families (6%) did not pass quality control, leaving 225 families (200 quartets, 25 trios) for analysis (Supplementary Data 1). Of the 200 quartets, 194 (97%) probands had a diagnosis of autism and 6 (3%) were diagnosed with ASD; the median non-verbal IQ was 84.

Exome capture, sequencing and variant prediction. Whole-blood DNA was enriched for exonic sequences through hybridization with a NimbleGen custom array ($N = 210$) or EZExomeV2.0 ($N = 718$). Captured DNA was sequenced using an Illumina GAIIx ($N = 592$) or HiSeq 2000 ($N = 336$). Short read sequences were aligned to hg18 with BWA[6], duplicate reads were removed and variants were predicted using SAMtools[7]. Data were normalized within families by only analysing bases with at least 20 unique reads in all family members. *De novo* predictions were made blinded to affected status using experimentally verified thresholds (Supplementary Information). All *de novo* variants were confirmed using Sanger sequencing blinded to affected status.

Gene annotation. Variants were analysed against RefSeq hg18 gene definitions; in genes with multiple isoforms the most severe outcome was chosen. All nonsense and canonical splice site variants were present in all RefSeq isoforms. A variant was listed as altering the splice site only if it disrupted canonical 2-base-pair acceptor (AG) or donor (GT) sites. Brain-expressed genes were identified from expression array analysis across 57 post-mortem brains (age 6 weeks post conception to 82 years) and multiple brain regions; 80% of RefSeq genes were included in this subset[24].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 September 2011; accepted 14 February 2012.
Published online 4 April 2012.

1. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
2. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
3. Sanders, S. J. *et al.* Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
4. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature Genet.* **43,** 585–589 (2011).
5. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
7. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
8. Meisler, M. H., O'Brien, J. E. & Sharkey, L. M. Sodium channel gene family: epilepsy mutations, gene interactions and modifier effects. *J. Physiol. (Lond.)* **588,** 1841–1848 (2010).
9. Kamiya, K. *et al.* A nonsense mutation of the sodium channel gene *SCN2A* in a patient with intractable epilepsy and mental decline. *J. Neurosci.* **24,** 2690–2698 (2004).
10. Ogiwara, I. *et al. De novo* mutations of voltage-gated sodium channel alphall gene *SCN2A* in intractable epilepsies. *Neurology* **73,** 1046–1053 (2009).
11. Weiss, L. A. *et al.* Sodium channels *SCN1A*, *SCN2A* and *SCN3A* in familial autism. *Mol. Psychiatry* **8,** 186–194 (2003).
12. Xu, B. *et al.* Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nature Genet.* **43,** 864–868 (2011).
13. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
14. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* **4,** 1073–1081 (2009).
15. Cooper, G. M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature Methods* **7,** 250–251 (2010).
16. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
17. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).
18. Abul-Husn, N. S. *et al.* Systems approach to explore components and interactions in the presynapse. *Proteomics* **9,** 3303–3315 (2009).
19. Bayés, A. *et al.* Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nature Neurosci.* **14,** 19–21 (2011).
20. Collins, M. O. *et al.* Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. *J. Neurochem.* **97** (suppl. 1), 16–23 (2006).
21. Girard, S. L. *et al.* Increased exonic de novo mutation rate in individuals with schizophrenia. *Nature Genet.* **43,** 860–863 (2011).
22. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
23. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* http://dx.doi.org/10.1038/nature10989 (this issue).
24. Kang, H. J. *et al.* Spatio-temporal transcriptome of the human brain. *Nature* **478,** 483–489 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to all of the families participating in the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection (SSC). This work was supported by a grant from the Simons Foundation. R.P.L. is an Investigator of the

Howard Hughes Medical Institute. We thank the SSC principal investigators A. L. Beaudet, R. Bernier, J. Constantino, E. H. Cook Jr, E. Fombonne, D. Geschwind, D. E. Grice, A. Klin, D. H. Ledbetter, C. Lord, C. L. Martin, D. M. Martin, R. Maxim, J. Miles, O. Ousley, B. Peterson, J. Piggot, C. Saulnier, M. W. State, W. Stone, J. S. Sutcliffe, C. A. Walsh and E. Wijsman and the coordinators and staff at the SSC sites for the recruitment and comprehensive assessment of simplex families; the SFARI staff, in particular M. Benedetti, for facilitating access to the SSC; Prometheus Research for phenotypic data management and Prometheus Research and the Rutgers University Cell and DNA repository for accessing biomaterials; the Yale Center of Genomic Analysis, in particular M. Mahajan, S. Umlauf, I. Tikhonova and A. Lopez, for generating sequencing data; T. Brooks-Boone, N. Wright-Davis and M. Wojciechowski for their help in administering the project at Yale; I. Hart for support; G. D. Fischbach, A. Packer, J. Spiro, M. Benedetti and M. Carlson for their suggestions throughout; and B. Neale and M. Daly for discussions regarding *de novo* variation. We also acknowledge T. Lehner and the Autism Sequencing Consortium for providing an opportunity for pre-publication data exchange among the participating groups.

Author Contributions S.J.S., M.T.M., R.P.L., M.G., D.H.G. and M.W.S. designed the study; M.T.M., A.R.G., J.M., M.R., A.G.E.-S., N.M.D., S.M., M.W., G.O., Y.S., P.E., R.M. and J.O. designed and performed high-throughput sequencing experiments and variant confirmations; S.J.S., M.C., K.B., R.B. and N.C. designed the exome-analysis bioinformatics pipeline; S.J.S., A.J.W., N.N.P., J.L.S., N.T., K.A.M., N.Š., K.R., D.H.G., B.D. and M.W.S. analysed the data; S.J.S., A.J.W., K.R., B.D. and M.W.S. wrote the paper; J.M., M.R., A.J.W., A.R.G., A.G.E.-S. and N.M.D. contributed equally to the study. All authors discussed the results and contributed to editing the manuscript.

Author Information Sequence data from this study is available through the NCBI Sequence Read Archive (accession number SRP010920.1). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.W.S. (matthew.state@yale.edu), B.D. (devlinbj@upmc.edu) or D.H.G. (dhg@mednet.ucla.edu).

METHODS

Sample selection. In total 238 families (928 individuals) were selected from the SSC on the basis of: male probands with autism, low non-verbal IQ (NVIQ), and discordant Social Responsiveness Scale (SRS) with sibling and parents ($N = 40$); female probands ($N = 46$); multiple unaffected siblings ($N = 28$); probands with known multigenic CNVs ($N = 15$); and random selection ($N = 109$). Thirteen families (6%) did not pass quality control (Supplementary Information) leaving 225 families (200 quartets, 25 trios) for analysis (Supplementary Data 1). Of the 200 quartets, 194 (97%) probands had a diagnosis of autism and 6 (3%) were diagnosed with ASD; the median NVIQ was 84. Three of these quartets have previously been reported as trios[4]; there is no overlap between the current sample and those presented in the companion article[23].

Exome capture, sequencing and variant prediction. Whole-blood DNA was enriched for exonic sequences (exome capture) through hybridization with a NimbleGen custom array ($N = 210$) or EZExomeV2.0 ($N = 718$). The captured DNA was sequenced using an Illumina GAIIx ($N = 592$) or HiSeq 2000 ($N = 336$). Short read sequences were aligned to hg18 with BWA[6], duplicate reads were removed and variants were predicted using SAMtools[7]. The data were normalized across each family by only analysing bases with at least 20 unique reads in all family members (Supplementary Information). *De novo* predictions were made blinded to affected status using experimentally verified thresholds (Supplementary Information). All *de novo* variants were confirmed using Sanger sequencing blinded to affected status.

Variant frequency. The allele frequency of a given variant in the offspring was determined by comparison with dbSNPv132 and 1,637 whole-exome controls including 400 parents. Variants were classified as: 'novel', if only a single allele was present in a parent and none were seen in dbSNP or the other control exomes; 'rare', if they did not meet the criteria for novel and were present in $<1\%$ of controls; and 'common', if they were present in $\geq 1\%$ of controls.

Gene annotation. Variants were analysed against the RefSeq hg18 gene definitions, a list that includes 18,933 genes. Where multiple isoforms gave varying results the most severe outcome was chosen. All nonsense and canonical splice site variants were checked manually and were present in all RefSeq isoforms. A variant was listed as altering the splice site only if it disrupted canonical 2-base-pair acceptor (AG) or donor (GT) sites.

Brain-expressed genes. A list of brain-expressed genes was obtained from expression array analysis across 57 post-mortem brains (age 6 weeks post conception to 82 years) and multiple brain regions[24]. Using these data, 14,363 (80%) of genes were classified as brain-expressed (Supplementary Information).

Rate of *de novo* SNVs. To allow an accurate comparison between the *de novo* burden in probands and siblings, the number of *de novo* SNVs found in each sample was divided by the number of bases analysed (that is, bases with ≥ 20 unique reads in all family members) to calculate a per-base rate of *de novo* SNVs. Rates are given in Table 1.

Simulation model. The likelihood of observing multiple independent *de novo* events of a given type for a given sample size in an ASD risk-conferring gene was modelled using gene size and GC content (derived from the full set of brain-expressed RefSeq genes) and the observed rate of brain-expressed *de novo* variants in probands and siblings. These values were then used to evaluate the number of genes contributing to ASD showing two or more variants of the specified type (Fig. 2); comparing this to the number of genes with similar events not carrying ASD risk gave the likelihood of the specified pattern demonstrating association with ASD. The simulation was run through 150,000 iterations across a range of samples sizes and multiple models of locus heterogeneity (Supplementary Information).

Severity scores. Severity scores were calculated for missense variants using web-based interfaces for PolyPhen2[13], SIFT[14] and GERP[15], using the default settings (Supplementary Information). PhyloP[16] and Grantham Score[17] were determined using an in-house annotated script. For nonsense/splice site variants the maximum score was assigned for Grantham, SIFT and PolyPhen2; for GERP and PhyloP, every possible coding base for the specific protein was scored and the highest value selected.

Pathway analysis. The list of brain-expressed genes with non-synonymous *de novo* SNVs was submitted to KEGG using the complete set of 14,363 brain-expressed genes as the background to prevent bias. For IPA the analysis was based on human nervous system pathways only, again to prevent bias. Otherwise default settings were used for both tools.

Protein–protein interactions. Genes with brain-expressed non-synonymous *de novo* variants in probands were submitted to the Disease Association Protein–protein Link Evaluator (DAPPLE)[22] using the default settings.

Comparing *de novo* SNV counts to gene lists. To assess whether non-synonymous *de novo* SNVs were enriched in particular gene sets, the chance of seeing a *de novo* variant in each gene on a given list was estimated based on the size and GC content of the gene. The observed number of *de novo* events was then assessed using the binomial distribution probability based on the total number of non-synonymous *de novo* variants in probands and the sum of probabilities for *de novo* events within these genes.

LETTER

doi:10.1038/nature11011

Patterns and rates of exonic *de novo* mutations in autism spectrum disorders

Benjamin M. Neale[1,2], Yan Kou[3,4], Li Liu[5], Avi Ma'ayan[3], Kaitlin E. Samocha[1,2], Aniko Sabo[6], Chiao-Feng Lin[7], Christine Stevens[2], Li-San Wang[7], Vladimir Makarov[4,8], Paz Polak[2,9], Seungtai Yoon[4,8], Jared Maguire[2], Emily L. Crawford[10], Nicholas G. Campbell[10], Evan T. Geller[7], Otto Valladares[7], Chad Schafer[5], Han Liu[11], Tuo Zhao[11], Guiqing Cai[4,8], Jayon Lihm[4,8], Ruth Dannenfelser[3], Omar Jabado[12], Zuleyma Peralta[12], Uma Nagaswamy[6], Donna Muzny[6], Jeffrey G. Reid[6], Irene Newsham[6], Yuanqing Wu[6], Lora Lewis[6], Yi Han[6], Benjamin F. Voight[2,13], Elaine Lim[1,2], Elizabeth Rossin[1,2], Andrew Kirby[1,2], Jason Flannick[2], Menachem Fromer[1,2], Khalid Shakir[2], Tim Fennell[2], Kiran Garimella[2], Eric Banks[2], Ryan Poplin[2], Stacey Gabriel[2], Mark DePristo[2], Jack R. Wimbish[14], Braden E. Boone[14], Shawn E. Levy[14], Catalina Betancur[15], Shamil Sunyaev[2,9], Eric Boerwinkle[6,16], Joseph D. Buxbaum[4,8,12,17], Edwin H. Cook Jr[18], Bernie Devlin[19], Richard A. Gibbs[6], Kathryn Roeder[5], Gerard D. Schellenberg[7], James S. Sutcliffe[10] & Mark J. Daly[1,2]

Autism spectrum disorders (ASD) are believed to have genetic and environmental origins, yet in only a modest fraction of individuals can specific causes be identified[1,2]. To identify further genetic risk factors, here we assess the role of *de novo* mutations in ASD by sequencing the exomes of ASD cases and their parents ($n = 175$ trios). Fewer than half of the cases (46.3%) carry a missense or nonsense *de novo* variant, and the overall rate of mutation is only modestly higher than the expected rate. In contrast, the proteins encoded by genes that harboured *de novo* missense or nonsense mutations showed a higher degree of connectivity among themselves and to previous ASD genes[3] as indexed by protein-protein interaction screens. The small increase in the rate of *de novo* events, when taken together with the protein interaction results, are consistent with an important but limited role for *de novo* point mutations in ASD, similar to that documented for *de novo* copy number variants. Genetic models incorporating these data indicate that most of the observed *de novo* events are unconnected to ASD; those that do confer risk are distributed across many genes and are incompletely penetrant (that is, not necessarily sufficient for disease). Our results support polygenic models in which spontaneous coding mutations in any of a large number of genes increases risk by 5- to 20-fold. Despite the challenge posed by such models, results from *de novo* events and a large parallel case–control study provide strong evidence in favour of *CHD8* and *KATNAL2* as genuine autism risk factors.

In spite of the substantial heritability, few genetic risk factors for ASD have been identified[1,2]. Copy number variants (CNVs), in particular *de novo* and large events spanning multiple genes, have been identified as conferring risk[4,5]. Although these CNVs provide important leads to underlying biology, they rarely implicate single genes, are rarely fully penetrant, and many confer risk to a broad range of conditions including intellectual disability, epilepsy and schizophrenia[6]. There are also documented instances of rare single nucleotide variants (SNVs) that are highly penetrant for ASD[3].

Large-scale genetic studies make clear that the origins of ASD risk are multifarious, and recent estimates based on CNV data put the number of independent risk loci in the hundreds[5]. Yet knowledge regarding specific risk-determining genes and the overall genetic architecture for ASD remains incomplete. Although new sequencing technologies provide a catalogue of most variation in the genome, the profound locus heterogeneity of ASD makes it challenging to distinguish variants that confer risk from the background noise of inconsequential SNVs. *De novo* variation, being less frequent and potentially more deleterious, could offer insights into risk-determining genes. Accordingly, we sought to evaluate carefully the observed rate and consequence of *de novo* point mutations in the exomes of ASD subjects.

We performed exome sequencing of 175 ASD probands and their parents across five centres with multiple protocols and validation techniques (Supplementary Information). We used a sensitive and specific analytical pipeline based on current best practices[7–9] to analyse all data and observed no heterogeneity of mutation rate across centres.

In the entire sample, we observed 161 coding region point mutations (101 missense, 50 silent and 10 nonsense), with an additional two conserved splice site (CSS) SNVs and six frameshift insertions/deletions (indels) validated and included in pathway analyses (Supplementary Table 1).

To determine whether the rate of coding region point mutations was elevated, we estimated the mutation rate in light of coverage and base context using two parallel approaches (Supplementary Information). On the basis of both models, the exome target should have a significantly increased (~30%) mutation rate compared to the genome. Conservatively, by assuming the low end of the estimated mutation rate from recent whole-genome data (1.2×10^{-8})[10], we estimate a mutation rate of 1.5×10^{-8} for the exome sequence captured here. The observed point mutation rate of 0.92 per exome is slightly but not significantly elevated versus expectation (Table 1) and is insensitive to adjustment for lower coverage regions (Supplementary Information). Indeed our rate is similar to that of ref. 11.

Per-family events were distributed exquisitely according to the Poisson distribution (Table 1), suggesting limited variation in the underlying rate of *de novo* mutation in ASD families. The relative rates

[1]Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, New York, New York 10029, USA. [4]Seaver Autism Center for Research and Treatment, Mount Sinai School of Medicine, New York, New York 10029, USA. [5]Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15232, USA. [6]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [7]Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [8]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA. [9]Division of Genetics, Department of Medicine Brigham & Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. [10]Vanderbilt Brain Institute, Departments of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt University, Nashville, Tennessee 37232, USA. [11]Biostatistics Department and Computer Science Department, Johns Hopkins University, Baltimore, Maryland 21205, USA. [12]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York 10029, USA. [13]Department of Pharmacology, University of Pennsylvania, Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA. [14]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA. [15]INSERM U952 and CNRS UMR 7224 and UPMC Univ Paris 06, 75005 Paris, France. [16]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77030, USA. [17]Friedman Brain Institute, Mount Sinai School of Medicine, New York, New York 10029, USA. [18]Department of Psychiatry, University of Illinois at Chicago, Chicago, Illinois 60608, USA. [19]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA.

Table 1 | Distribution of events per family

Events per family	All ASD trios		Random mut. exp.‡
	Exon DN SNVs*	Exp.†	
0	71	69.7	73.2
1	62	64.2	63.8
2	28	29.5	27.8
3	10	9.1	8.1
4	2	2.1	1.8
5	1	0.4	0.3
Mean		0.920	0.871

* Exon DN SNVs include all single nucleotide variants in coding sequence but excludes indels and intronic variants.
† The expected distribution of number of trios with a given event count as determined by the Poisson.
‡ Random mut. exp. is the expectation for 175 trios based on the sequence-context mutation rate model M1 (Supplementary Information) based on the count of the number of trios that have at least 10× coverage.

of 'functional' (missense, nonsense, CSS and read-through) versus silent changes did not deviate from expectation (Table 2). We did, however, observe ten nonsense mutations (6.2%), which exceeded expectation (3.3%) (one-tailed $P = 0.04$; Supplementary Information).

We examined missense mutations using PolyPhen-2 scores[12] to measure severity, as some missense variants can severely affect function[13]. These scores showed no deviation from random expectation. The observed PolyPhen-2 scores clearly deviate from standing variation in the parents (Table 2), but such variation, even the rarest category, has survived selective pressure and so is inappropriate for comparison to *de novo* events.

We observed three genes with two *de novo* mutations: *BRCA2* (two missense), *FAT1* (two missense) and *KCNMA1* (one missense, one silent). A gene with two or more non-synonymous *de novo* hits across a panel of trios might indicate strong candidacy. However, simulations (Supplementary Information) show that two such hits are inadequate to define a gene as a conclusive risk factor given the number of observed events in the study.

From analyses of secondary phenotypes (Supplementary Tables 2 and 3), the most striking result is that paternal and maternal age, themselves highly correlated ($r^2 = 0.679$, P-value <0.0001), each strongly predicts the number of *de novo* events per offspring (paternal age, $P = 0.0013$; maternal age, $P = 0.000365$), consistent with aggregating mutations in germ cells in the paternal line[14]. Consistent with a liability threshold model, there is an increased rate of *de novo* mutation in female versus male cases (1.214 for females versus 0.914 for males); however, the difference is not significant, owing to limited sample size. Considering phenotypic correlates, we observed no rate difference between subjects with strict autism versus those with a broader ASD classification, between positive and negative family history, or any significant effect of *de novo* mutation on verbal, non-verbal or full-scale IQ (Supplementary Table 3).

Given that hundreds of loci are apparently involved in autism[5] and *de novo* mutations therein affect ASD risk, we modelled different numbers of risk genes and penetrances (Supplementary Information) and show that a model of hundreds of genes with high penetrance mutations is excluded by our data; however, more modest contributions of *de novo* variants are not. For example, up to 20% of cases carrying a *de novo* event conferring a 10- or 20-fold increased risk is consistent with these data (Supplementary Table 4). Thus, our data are consistent with either chance mutation or a modest role for *de novo* mutations on risk. Importantly, a single deleterious event is unlikely to fully explain disease in a patient.

Table 2 | Rates of mutation annotation given variant type

Type of *de novo* mutation	*De novo* (%)*	Random *de novo* (%)	Singletons (%)†	Doubletons (%)†	≥3 (%)†
Missense	62.7	66.1	59.5	55.4	48.8
Nonsense	6.2	3.3	1.2	0.8	0.4
Synonymous	31.1	30.6	39.3	43.8	50.8
PolyPhen-2 missense classification					
Benign	35.0	35.9	46.6	51.3	63.4
Possibly damaging	21.0	18.9	18.8	17.7	15.1
Probably damaging	44.0	45.2	34.7	31.0	21.4

* All indels and failing variants were removed.
† Singletons, doubletons and ≥3 (copies) are only those variants called in 192 parents.

We therefore posed two questions of the group of genes harbouring *de novo* functional mutations: do the protein products of these genes interact with each other more than expected, and are they unusually enriched in, or connected to, previous curated lists of ASD-implicated genes? Using an *in silico* approach (DAPPLE)[15], the protein–protein connectivity defined by InWeb[16] in the set of 113 genes harbouring functional *de novo* mutations was evaluated. These analyses (Fig. 1) showed significantly greater connectivity among the *de novo* identified proteins than would be expected by chance ($P < 0.001$) (Supplementary Information).

Querying previously defined, manually curated lists of genes[3] associated with high risk for ASD with or without intellectual disability (Supplementary Table 5), and high-risk intellectual disability genes (Supplementary Table 6), we asked whether there was significant enrichment for *de novo* mutations in these genes. Five genes with functional *de novo* events were previously associated with ASD and/or intellectual disability (*STXBP1*, *MEF2C*, *KIRREL3*, *RELN* and *TUBA1A*); for four of these genes (all but *RELN*) the previous evidence indicated autosomal dominant inheritance.

We then assessed the average distance (D_i, Supplementary Fig. 2) of the *de novo* coding variants in brain-expressed genes (see supplement) to the ASD/intellectual disability list using a protein–protein interaction background network. To enhance power, data from a companion study[11] were used, including the observed silent *de novo* variants and *de novo* variants in unaffected siblings as comparators. The average distance for non-synonymous variants was significantly smaller for the case set than the comparator set (3.66 ± 0.42 versus 3.78 ± 0.59; permutation $P = 0.033$) (Supplementary Fig. 3). Much of this signal comes from 31 synaptic genes identified by three large-scale synaptic proteomic studies ($D_i = 3.47 \pm 0.46$ versus 3.57 ± 0.60; permutation $P = 0.084$) (Fig. 2; see also Supplementary Fig. 4 for the complete data). Taken in total, these independent gene set analyses, along with the modest enrichment of *de novo* variants over background rates in

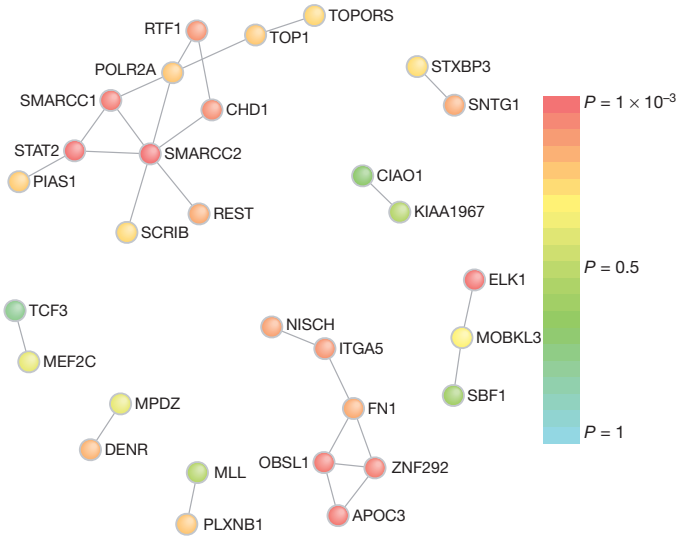


Figure 1 | Protein–protein interaction for genes with an observed functional *de novo* event. Direct protein connections from InWeb, restricting to genes harbouring *de novo* mutations for DAPPLE analysis. Two extensive networks are identified: the first is centred on SMARCC2 with 12 connections across 11 genes; the second is centred on FN1 with 7 connections across 6 genes. The P value for each gene having as many connections as those observed is indicated by node colour.

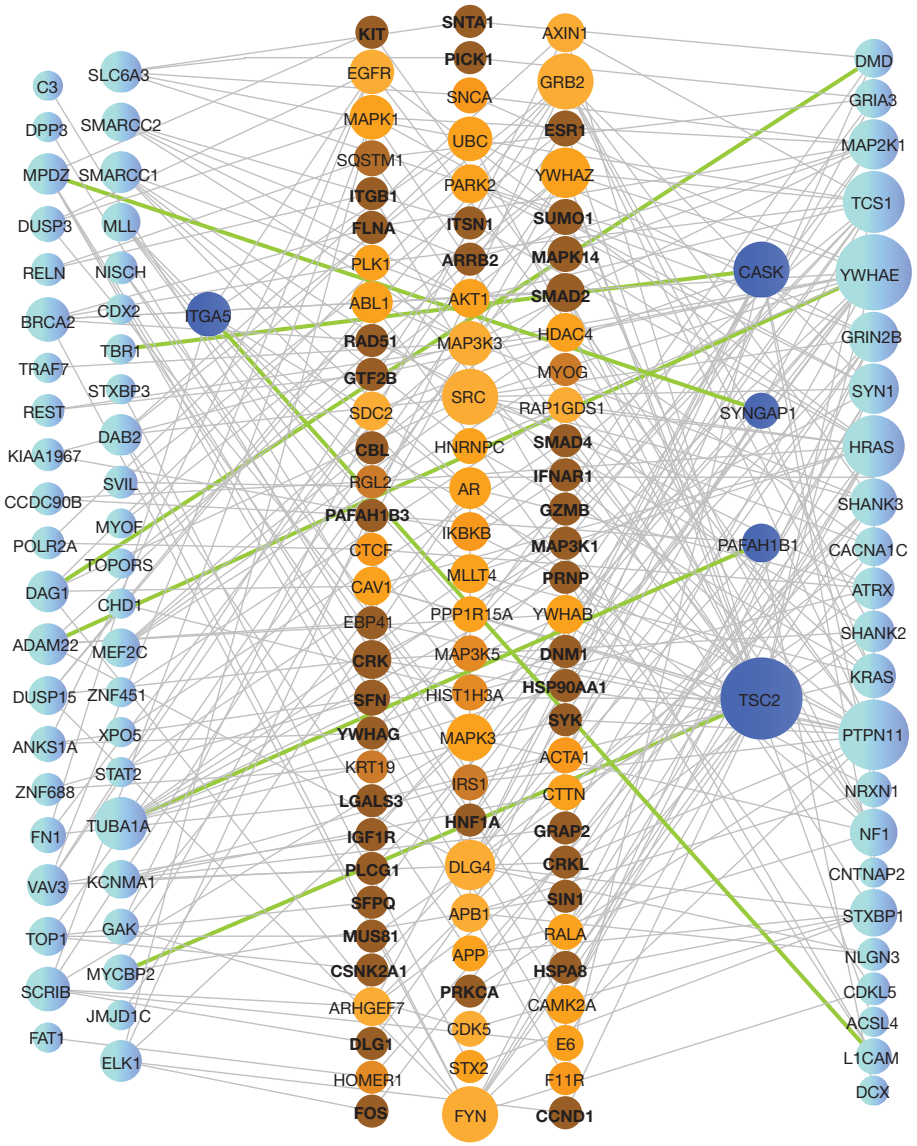


Figure 2 | Direct and indirect protein–protein interaction for genes with a functional *de novo* event and previous ASD genes. PPI network analysis for *de novo* variants and 31 previous synaptic ASD genes (see Supplementary Information). Nodes are sized based on connectivity. Genes harbouring *de novo* variants (left) and previous ASD genes (right) are coloured blue, with dark blue nodes representing genes that belong to one of these lists and are also intermediate proteins. Intermediate proteins (centre) are coloured in shades of orange based on a *P* value computed using a proportion test, where a darker colour represents a lower *P* value. Green edges represent direct connections between genes harbouring *de novo* variants (left) and previous ASD genes. All other edges, connecting to intermediate proteins, are shown in grey.

ASD, indicate that a proportion of the *de novo* events observed in this study probably contribute to autism risk.

Using whole-exome sequencing of autism trios, we demonstrate a rate, functional distribution and predicted impact of *de novo* mutation largely consistent with chance mutational processes governed by sequence context. This lack of significant deviation from random mutational processes indicates a more limited role for the contribution of *de novo* mutations to ASD pathogenesis than has previously been suggested[17], and specifically highlights the fact that observing a single *de novo* mutation, even an apparently 'severe' loss-of-function allele, is insufficient to implicate a gene as a risk factor. Yet the pathway analyses presented here assert that the overall set of genes hit with functional *de novo* mutations is not random and that these genes are biologically related to each other and to previously identified ASD/intellectual disability candidate genes. Modelling the *de novo* mutational process under a range of genetic models reveals that some models are inconsistent with the observed data—for example, 100 rare, fully penetrant Mendelian genes similar to Rett's syndrome—whereas others are not inconsistent, such as spontaneous 'functional' mutation in hundreds of genes that would increase risk by 10- or 20-fold (Supplementary Table 4). Models that fit the data are consistent with the relative risks estimated for most *de novo* CNVs[5] and suggest that *de novo* SNVs, like most CNVs, often combine with other risk factors rather than fully cause disease. Furthermore, these models indicate that *de novo* SNV events will probably explain <5% of the overall variance in autism risk (Supplementary Table 4).

Considering the two companion papers[11,18], 18 genes with two functional *de novo* mutations are observed in the complete data. Using simulations, 11.91 genes on average harbour functional mutations by chance (Supplementary Table 7). Thus, a set of 18 genes with two or more hits is not quite significant ($P = 0.063$). Matching loss-of-function variants, however, at *SCN2A*, *KATNAL2* and *CHD8* (Supplementary Table 7) are unlikely to occur by chance because of the expected very low rate of *de novo* nonsense, splice and frameshift variants. We evaluated these strong candidates further using exome sequencing on 935 cases and 870 controls, and at both *KATNAL2* and

CHD8 three additional loss-of-function mutations were observed in cases with none in controls. No additional loss-of-function mutations were seen at *SCN2A* in the case–control data, but a new splice site *de novo* event has been validated in an additional autism case while this paper was in press, strengthening the evidence for this gene as relevant to autism. Using data from more than 5,000 individuals in the NHLBI Exome Variant Server (http://evs.gs.washington.edu/EVS/) as additional controls, three loss-of-function mutations were seen in *KATNAL2* but none in *CHD8*, making the additional observation of three *CHD8* loss-of-function mutations in our cases significant evidence ($P < 0.01$) of this being a genuine autism susceptibility gene. Not all genes with double hits are nearly so promising (Supplementary Information and Supplementary Tables 8 and 9), supporting the estimate above that most of such observations are simply chance events. Overall, these data underscore the challenge of establishing individual genes as conclusive risk factors for ASD, a challenge that will require larger sample sizes and deeper analytical integration with inherited variation.

METHODS SUMMARY

We ascertained probands using the Autism Diagnostic Interview-Revised (ADI-R), the Autism Diagnostic Observation Schedule-Generic (ADOS) and the DSM-IV diagnosis of a pervasive developmental disorder. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects that were not assessed with the ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

For 175 trios, we performed exome capture and sequencing using either the Agilent 38Mb SureSelect v2 ($n = 118$), the NimbleGen Seq Cap EZ SR v2 ($n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

All sequence data were processed with Picard (http://picard.sourceforge.net/), which recalibrates quality scores and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. Putative *de novo* mutations were identified restricting to sites passing standard filters and both parents were homozygous for the reference sequence and the offspring was heterozygous, and each genotype call was made confidently (see Supplementary Information).

All putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (71 trios) or by using Sequenom MALDI-TOF (104 trios). All events were annotated using RefSeq hg19.

We modelled a Poisson process consistent with the mutation model and observed data. We varied the fraction of genes that influence risk, the probability of a functional variant, and the penetrance of said events.

We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes × 2 sites), but we also performed mega- and meta-analyses (23 genes × 2 methods).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 13 September 2011; accepted 6 March 2012.
Published online 4 April 2012.

1. Lichtenstein, P., Carlstrom, E., Rastam, M., Gillberg, C. & Anckarsater, H. The genetics of autism spectrum disorders and related neuropsychiatric disorders in childhood. *Am. J. Psychiatry* **167,** 1357–1363 (2010).
2. Hallmayer, J. *et al.* Genetic heritability and shared environmental factors among twin pairs with autism. *Arch. Gen. Psychiatry* **68,** 1095–1102 (2011).
3. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
4. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466,** 368–372 (2010).
5. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
6. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25,** 528–535 (2009).
7. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
8. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
9. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
10. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43,** 712–714 (2011).
11. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* http://dx.doi.org/10.1038/nature10945 (this issue).
12. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).
13. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80,** 727–739 (2007).
14. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1,** 40–47 (2000).
15. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
16. Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA* **105,** 20870–20875 (2008).
17. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
18. O'Roak, B. J. *et al.* Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature* http://dx.doi.org/10.1038/nature10989 (this issue).
19. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89,** 82–93 (2011).
20. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* **7,** e1001322 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was directly supported by NIH grants R01MH089208 (M.J.D.), R01 MH089025 (J.D.B.), R01 MH089004 (G.D.S.), R01MH089175 (R.A.G.) and R01 MH089482 (J.S.S.), and supported in part by NIH grants P50 HD055751 (E.H.C.), RO1 MH057881 (B.D.) and R01 MH061009 (J.S.S.). Y.K., G.C. and S.Y. are Seaver Fellows, supported by the Seaver Foundation. We thank T. Lehner, A. Felsenfeld and P. Bender for their support and contribution to the project. We thank S. Sanders and M. State for discussions on the interpretation of *de novo* events. We thank D. Reich for comments on the abstract and message of the manuscript. We thank E. Lander and D. Altshuler for comments on the manuscript. We acknowledge the assistance of M. Potter, A. McGrew and G. Crockett without whom these studies would not be possible, and Center for Human Genetics Research resources: Computational Genomics Core, Genetic Studies Ascertainment Core and DNA Resources core, supported in part by NIH NCRR grant UL1 RR024975, and the Vanderbilt Kennedy Center for Research on Human Development (P30 HD015052). This work was supported in part by R01MH084676 (S.S.). We acknowledge the clinicians and organizations that contributed to samples used in this study and the particular support of the Mount Sinai School of Medicine, University of Illinois-Chicago, Vanderbilt University, the Autism Genetics Resource Exchange and the institutions of the Boston Autism Consortium. We acknowledge A. Estes and G. Dawson for patient collection/characterization. We acknowledge partial support from U54 HG003273 (R.A.G.) and U54 HG003067 (E. Lander). J.D.B., B.D., M.J.D., R.A.G., A.S., G.D.S. and J.S.S. are lead investigators in the Autism Sequencing Consortium (ASC). The ASC is comprised of groups sharing massively parallel sequencing data in autism. Finally, we are grateful to the many families, without whose participation this project would not have been possible.

Author Contributions Laboratory work: A.S., C.St., G.C., O.J., Z.P., J.D.B., D.M., I.N., Y.W., L.L., Y.H., S.G., E.L.C., N.G.C. and E.T.G. Data processing: B.M.N., K.E.S., E.L., A.K., J.F., M.F., K.S., T.F., K.G., E.Ba., R.P., M.DeP., S.G., S.Y., V.M., J.L., J.D.B., A.S., C.St., U.N., J.G.R., J.R.W., B.E.B., S.E.L., C.F.L., L.S.W. and O.V. Statistical analysis: B.M.N., L.L., K.E.S., C.Sh., B.F.V., J.M., E.R., S.S., P.P., Y.K., A.M., R.D., C.-F.L., L.-S.W., H.L., T.Z., E.Bo., R.A.G., J.D.B., C.B., E.H.C., J.S.S., G.D.S., B.D., K.R. and M.J.D. Principal Investigators/study design: E.Bo., R.A.G., E.H.C., J.D.B., K.R., B.D., G.D.S., J.S.S. and M.J.D. Y.K., L.L., A.M., K.E.S., A.S. and C.-F.L. contributed equally to this work. E.Bo., J.D.B., E.H.C., B.D., R.A.G., K.R., G.D.S., J.S.S. and M.J.D. are lead investigators of the ARRA Autism Sequencing Collaboration.

Author Information Data included in this manuscript have been deposited at dbGaP under accession number phs000298.v1.p1 and is available for download at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000298.v1.p1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.J.D. (mjdaly@atgu.mgh.harvard.edu), J.D.B. (joseph.buxbaum@mssm.edu) or K.R. (kathryn.roeder@gmail.com).

METHODS

Phenotype assessment. Affected probands were assessed by research-reliable research personnel using Autism Diagnostic Interview-Revised (ADI-R), and the Autism Diagnostic Observation Schedule-Generic (ADOS) and DSM-IV diagnosis of a pervasive developmental disorder was made by a clinician. All probands met criteria for autism on the ADI-R and either autism or ASD on the ADOS, except for the three subjects from AGRE that were not assessed with the ADOS. In all, 85% of probands were classified with autism on both the ADI-R and ADOS. All subjects provided informed consent and the research was approved by institutional human subjects boards.

Exome sequencing, variant identification and *de novo* detection. Exome capture and sequencing was performed at each site using similar methods. Exons were captured using the Agilent 38 Mb SureSelect v2 (University of Pennsylvania and Broad Institute $n = 118$), the NimbleGen Seq Cap EZ SR v2 (Mt Sinai School of Medicine, Vanderbilt University $n = 51$), or NimbleGen VCRome 2.1 (Baylor $n = 6$). After capture, another round of LM-PCR was performed to increase the quantity of DNA available for sequencing. All libraries were sequenced using an IlluminaHiSeq2000.

Sequence processing and variant calling was performed using a similar computational workflow at all sites. Data were processed with Picard (http://picard.sourceforge.net/), which uses base quality-score recalibration and local realignment at known indels[8] and BWA[7] for mapping reads to hg19. SNPs were called using GATK[8,9] for all trios jointly. The variable sites that we have considered in analysis are restricted to those that pass GATK standard filters. From this set of variants, we identified putative *de novo* mutations as sites where both parents were homozygous for the reference sequence and the offspring was heterozygous and each genotype call was made confidently (see Supplementary Information).

Validation of *de novo* events. Putative *de novo* events were validated by sequencing the carrier and both parents using Sanger sequencing methods (University of Pennsylvania, Mt Sinai School of Medicine, Vanderbilt University, Baylor Medical College) or by Sequenom MALDI-TOF genotyping of trios (Broad).

Gene annotation. All identified mutations were then annotated using RefSeq hg19. The functional impact of variants was assessed for all isoforms of each gene, with the most severe annotation taking priority. Splice site variants were identified as occurring within two base pairs of any intron/exon boundary.

Expectation of *de novo* mutation calculation. To calculate the expected *de novo* rate, we assessed the mutability of all possible trinucleotide contexts in the intergenic region of the human genome for variation in two fashions: fixed genomic differences compared to chimpanzee and baboon[12] and variation identified from the 1,000 Genomes project. The overall mutation rate for the exome was then determined by summing the probability of mutation for all bases in the exome that were captured successfully. We also determined the probability of each class functional mutation by summing the annotated variants.

Pathway analyses. We applied DAPPLE[15], which uses the InWeb database[16], to determine whether there is excess protein–protein interaction across the genes hit by a functional *de novo* event. We also assessed whether these genes were more closely connected to a list of ASD genes[3].

Modelling *de novo* events. We modelled a Poisson process consistent with the expected distribution defined by the mutation model and with the observed data. We varied the fraction of genes that influence risk, the probability a variant in a gene would be functional, and the penetrance of functional *de novo* events. We also simulated a random set of *de novo* events to estimate the probability of hitting a gene multiple times.

Association analysis. We performed association tests using SKAT[19], a generalization of C-alpha[20]. Our primary analyses treat case–control data generated at Baylor and Broad sequencing centres separately (23 genes $\times$ 2 sites), but we also performed mega- and meta-analyses (23 genes $\times$ 2 methods).

LETTER

doi:10.1038/nature10989

Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations

Brian J. O'Roak[1], Laura Vives[1], Santhosh Girirajan[1], Emre Karakoc[1], Niklas Krumm[1], Bradley P. Coe[1], Roie Levy[1], Arthur Ko[1], Choli Lee[1], Joshua D. Smith[1], Emily H. Turner[1], Ian B. Stanaway[1], Benjamin Vernot[1], Maika Malig[1], Carl Baker[1], Beau Reilly[2], Joshua M. Akey[1], Elhanan Borenstein[1,3,4], Mark J. Rieder[1], Deborah A. Nickerson[1], Raphael Bernier[2], Jay Shendure[1] & Evan E. Eichler[1,5]

It is well established that autism spectrum disorders (ASD) have a strong genetic component; however, for at least 70% of cases, the underlying genetic cause is unknown[1]. Under the hypothesis that *de novo* mutations underlie a substantial fraction of the risk for developing ASD in families with no previous history of ASD or related phenotypes—so-called sporadic or simplex families[2,3]—we sequenced all coding regions of the genome (the exome) for parent–child trios exhibiting sporadic ASD, including 189 new trios and 20 that were previously reported[4]. Additionally, we also sequenced the exomes of 50 unaffected siblings corresponding to these new ($n = 31$) and previously reported trios ($n = 19$)[4], for a total of 677 individual exomes from 209 families. Here we show that *de novo* point mutations are overwhelmingly paternal in origin (4:1 bias) and positively correlated with paternal age, consistent with the modest increased risk for children of older fathers to develop ASD[5]. Moreover, 39% (49 of 126) of the most severe or disruptive *de novo* mutations map to a highly interconnected β-catenin/chromatin remodelling protein network ranked significantly for autism candidate genes. In proband exomes, recurrent protein-altering mutations were observed in two genes: *CHD8* and *NTNG1*. Mutation screening of six candidate genes in 1,703 ASD probands identified additional *de novo*, protein-altering mutations in *GRIN2B*, *LAMC3* and *SCN1A*. Combined with copy number variant (CNV) data, these results indicate extreme locus heterogeneity but also provide a target for future discovery, diagnostics and therapeutics.

We selected 189 autism trios from the Simons Simplex Collection (SSC)[6], which included males significantly impaired with autism and intellectual disability ($n = 47$), a female sample set ($n = 56$) of which 26 were cognitively impaired, and samples chosen at random from the remaining males in the collection ($n = 86$) (Supplementary Table 1 and Supplementary Fig. 1). In general, we excluded samples known to carry large *de novo* CNVs[2]. Exome sequencing was performed as described previously[4], but with an expanded target definition (see Methods). We achieved sufficient coverage for both parents and child to call genotypes for, on average, 29.5 megabases (Mb) of haploid exome coding sequence (Supplementary Table 1). In addition, we performed copy number analysis on 122 of these families, using a combination of the exome data, array comparative genomic hybridization (CGH), and genotyping arrays, thereby providing a more comprehensive view of rare variation.

In the 189 new probands, we validated 248 *de novo* events, 225 single nucleotide variants (SNVs), 17 small insertions/deletions (indels), and six CNVs (Supplementary Table 2). These included 181 non-synonymous changes, of which 120 were classified as severe based on sequence conservation and/or biochemical properties (Methods and Supplementary Table 3). The observed point mutation rate in coding sequence was ~1.3 events per trio or 2.17×10^{-8} per base per generation, in close agreement with our previous observations[4], yet in general, higher than previous studies, indicating increased sensitivity (Supplementary Table 2 and Supplementary Table 4)[7]. We also observed complex classes of *de novo* mutation including: five cases of multiple mutations in close proximity; two events consistent with paternal germline mosaicism (that is, where both siblings contained a *de novo* event observed in neither parent); and nine events showing a weak minor allele profile consistent with somatic mosaicism (Supplementary Table 3 and Supplementary Figs 2 and 3).

Of the severe *de novo* events, 28% (33 of 120) are predicted to truncate the protein. The distribution of synonymous, missense and nonsense changes corresponds well with a random mutation model[7] (Supplementary Fig. 4 and Supplementary Table 2). However, the difference in nonsense rates between *de novo* and rare singleton events (not present in 1,779 other exomes) is striking (4:1) and suggests strong selection against new nonsense events (Fisher's exact test, $P < 0.0001$). In contrast with a recent report[8], we find no significant difference in mutation rate between affected and unaffected individuals; however, we do observe a trend towards increased non-synonymous rates in probands, consistent with the findings of ref. 9 (Supplementary Tables 1 and 2).

Given the association of ASD with increased paternal age[5] and our previous observations[4], we used molecular cloning, read-pair information, and obligate carrier status to identify informative markers linked to 51 *de novo* events and observed a marked paternal bias (41:10; binomial $P < 1.4 \times 10^{-5}$; Fig. 1a and Supplementary Tables 3 and 5). This provides strong direct evidence that the germline mutation rate in protein-coding regions is, on average, substantially higher in males. A similar finding was recently reported for *de novo* CNVs[10]. In addition, we observe that the number of *de novo* events is positively correlated with increasing paternal age (Spearman's rank correlation = 0.19; $P < 0.008$; Fig. 1b). Together, these observations are consistent with the hypothesis that the modest increased risk for children of older fathers to develop ASD[5] is the result of an increased mutation rate.

Using sequence read-depth methods in 122 of the 189 families, we scanned ASD probands for either *de novo* CNVs or rare (<1% of controls), inherited CNVs. Individual events were validated by either array CGH or genotyping array (see Methods). We identified 76 events in 53 individuals, including six *de novo* (median size 467 kilobases (kb)) and 70 inherited (median size 155 kb) CNVs (Supplementary Table 6). These include disruptions of *EHMT1* (Kleefstra's syndrome, Online Mendelian Inheritance in Man (OMIM) accession 610253), *CNTNAP4* (reported in children with developmental delay and autism[11]) and the 16p11.2 duplication (OMIM 611913) associated with developmental delay, bipolar disorder and schizophrenia.

We performed a multivariate analysis on non-verbal IQ (NVIQ), verbal IQ (VIQ) and the load of 'extreme' *de novo* mutations—where extreme is defined as point mutations that truncate proteins, intersect

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA. [2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, Washington 98195, USA. [3]Department of Computer Science and Engineering, University of Washington, Seattle, Washington 98195, USA. [4]Santa Fe Institute, Santa Fe, New Mexico 87501, USA. [5]Howard Hughes Medical Institute, Seattle, Washington 98195, USA.

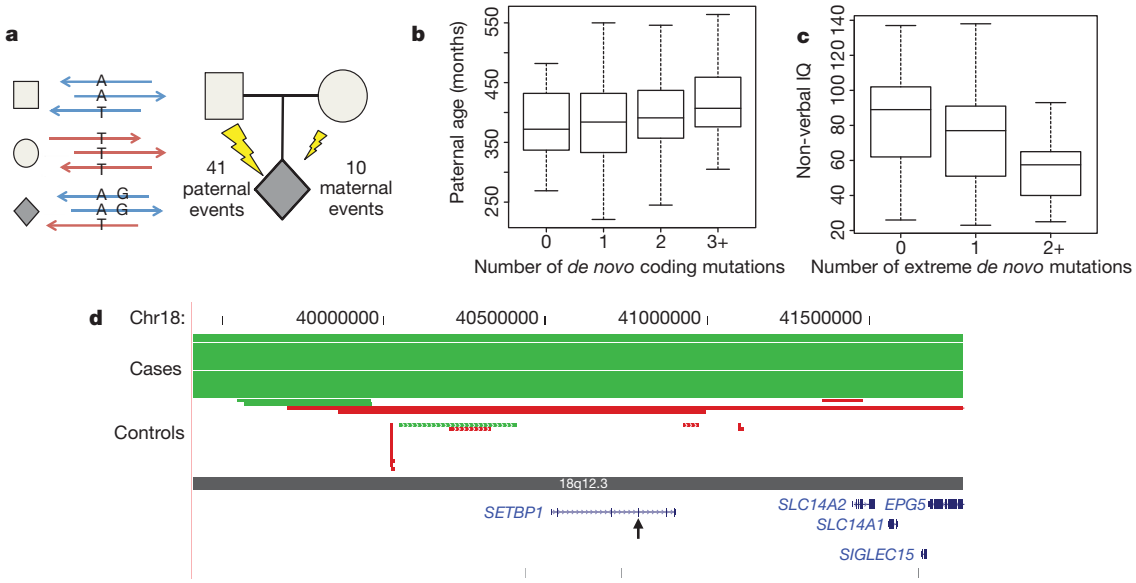


Figure 1 | *De novo* mutation events in autism spectrum disorder. **a**, Haplotype phasing using informative markers shows a strong parent-of-origin bias with 41 of 51 *de novo* events occurring on the paternally inherited haplotype. Arrows represent sequence reads from paternal (blue) or maternal (red) haplotypes. **b**, **c**, Box and whisker plots for 189 SSC probands. **b**, The paternal estimated age at conception versus the number of observed *de novo* point mutations (0, $n = 53$; 1, $n = 65$; 2, $n = 44$; 3+, $n = 27$). **c**, Decreased non-verbal IQ is significantly associated with an increasing number of extreme mutation events (0, $n = 138$; 1, $n = 41$; 2+, $n = 10$), both with and without CNVs (Supplementary Discussion). **d**, Browser images showing CNVs identified in the del(18)(q12.2q21.1) syndrome region. The truncating point mutation in *SETBP1* occurs within the critical region, identifying the likely causative locus. Each red (deletion) and green (duplication) line represents an identified CNV in cases (solid lines) versus controls (dashed lines), with arrowheads showing point mutation.

Mendelian or ASD loci ($n = 57$), or *de novo* CNVs that intersect genes ($n = 5$) (Fig. 1c and Supplementary Discussion). NVIQ, but not VIQ, decreased significantly ($P < 0.01$) with increased number of events. Covariant analysis of the samples with CNV data showed that this finding was strengthened, but not exclusively driven, by the presence of either *de novo* or rare CNVs (Supplementary Fig. 5).

Among the *de novo* events, we identified 62 top ASD risk contributing mutations based on the deleteriousness of the mutations, functional evidence, or previous studies (Table 1). Probands with these mutations spanned the range of IQ scores, with only a modest non-significant trend towards individual's co-morbid with intellectual disability (Supplementary Figs 1 and 6). We observed recurrent, protein-disruptive mutations in two genes: *NTNG1* (netrin G1) and *CHD8* (chromodomain helicase DNA binding protein 8). Given their locus-specific mutation rates, the probability of identifying two independent mutations in our sample set is low (uncorrected, *NTNG1*: $P < 1.2 \times 10^{-6}$; *CHD8*: $P < 6.9 \times 10^{-5}$) (Supplementary Fig. 7, Supplementary Table 8 and Methods). *NTNG1* is a strong biological candidate given its role in laminar organization of dendrites and axonal guidance[12] and was also reported as being disrupted by a *de novo* translocation in a child with Rett's syndrome, without *MECP2* mutation[13]. Both *de novo* mutations identified here are missense (p.Tyr23Cys and p.Thr135Ile) at highly conserved positions predicted to disrupt protein function, although there is evidence of mosaicism for the former mutation (Supplementary Table 3).

CHD8 has not previously been associated with ASD and codes for an ATP-dependent chromatin-remodelling factor that has a significant role in the regulation of both β-catenin and p53 signalling[14,15]. We also identified *de novo* missense variants in *CHD3* as well as *CHD7* (CHARGE syndrome, OMIM 214800), a known binding partner of CHD8 (ref. 16). ASD has been found in as many as two-thirds of children with CHARGE, indicating that *CHD7* may contribute to an ASD syndromic subtype[17].

We identified 30 protein-altering *de novo* events intersecting with Mendelian disease loci (Supplementary Table 3) as well as inherited hemizygous mutations of clinical significance (Supplementary Table 9). The *de novo* mutations included truncating events in syndromic intellectual disability genes (*MBD5* (mental retardation, autosomal dominant 1, OMIM 156200), *RPS6KA3* (Coffin–Lowry syndrome, OMIM 303600) and *DYRK1A* (the Down's syndrome candidate gene, OMIM 600855)), and missense variants in loci associated with syndromic ASD, including *CHD7*, *PTEN* (macrocephaly/autism syndrome, OMIM 605309) and *TSC2* (tuberous sclerosis complex, OMIM 613254). Notably, *DYRK1A* is a highly conserved gene mapping to the Down's syndrome critical region (Supplementary Fig. 8). The proband here (13890) is severely cognitively impaired and microcephalic, consistent with previous studies of *DYRK1A* haploinsufficiency in both patients and mouse models[18].

Twenty-one of the non-synonymous *de novo* mutations map to CNV regions recurrently identified in children with developmental delay and ASD (Supplementary Table 10), such as *MBD5* (2q23.1 deletion syndrome), *SYNRG* (17q12 deletion syndrome) and *POLRMT* (19p13.3 deletion)[19]. There is also considerable overlap with genes disrupted by single *de novo* CNVs in children with ASD (for example, *NLGN1* and *ARID1B*; Supplementary Table 11). Given the prior probability that these loci underlie genomic disorders, the disruptive *de novo* SNVs and small indels may be pinpointing the possible major effect locus for ASD-related features. For example, we identified a complex *de novo* mutation resulting in truncation of *SETBP1* (SET binding protein 1), one of five genes in the critical region for del(18)(q12.2q21.1) syndrome (Fig. 1d), which is characterized by hypotonia, expressive language delay, short stature and behavioural problems[20]. Recurrent *de novo* missense mutations at *SETBP1* were recently reported to be causative for a distinct phenotype, Schinzel–Giedion syndrome, probably through a gain-of-function mechanism[21], indicating diverse phenotypic outcomes at this locus depending on mutation mechanism.

Several of the mutated genes encode proteins that directly interact, suggesting a common biological pathway. From our full list of genes carrying truncating or severe missense mutations (126 events from all 209 families), we generated a protein–protein interaction (PPI) network based on a database of physical interactions (Supplementary Table 12)[22]. We found 39% (49 of 126) of the genes mapped to a highly

Table 1 | Top *de novo* ASD risk contributing mutations

Proband	NVIQ	Candidate gene	Amino acid change
12225.p1	89	*ABCA2*	p.Val1845Met
11653.p1	44	*ADCY5*	p.Arg603Cys
12130.p1	55	*ADNP*	Frameshift indel
11224.p1	112	*AP3B2*	p.Arg435His
13447.p1	51	*ARID1B*	Frameshift indel
13415.p1	48	*BRSK2*	3n indel
14292.p1	49	BRWD1	Frameshift indel
11872.p1	65	*CACNA1D*	p.Ala769Gly
11773.p1	50	*CACNA1E*	p.Gly1209Ser
13606.p1	60	*CDC42BPB*	p.Arg764TERM
12086.p1	108	*CDH5*	p.Arg545Trp
12630.p1	115	*CHD3*	p.Arg1818Trp
13733.p1	68	*CHD7*	p.Gly996Ser
13844.p1	34	*CHD8*	p.Gln959TERM
12752.p1	93	*CHD8*	Frameshift indel
13415.p1	48	*CNOT4*	p.Asp48Asn
12703.p1	58	*CTNNB1*	p.Thr551Met
11452.p1	80	*CUL3*	p.Glu246TERM
11571.p1	94	*CUL5*	p.Val355Ile
13890.p1	42	*DYRK1A*	Splice site
12741.p1	87	*EHD2*	p.Arg167Cys
11629.p1	67	*FBXO10*	p.Glu54Lys
13629.p1	63	*GPS1*	p.Arg492Gln
13757.p1	91	*GRINL1A*	3n indel
11184.p1	94	*HDGFRP2*	p.Glu83Lys
11610.p1	138	*HDLBP*	p.Ala639Ser
11872.p1	65	*KATNAL2*	Splice site
12346.p1	77	*MBD5*	Frameshift indel
11947.p1	33	*MDM2*	p.Glu433Lys/p.Trp160TERM
11148.p1	82	*MLL3*	p.Tyr4691TERM
12157.p1	91	*NLGN1*	p.His795Tyr
11193.p1	138	*NOTCH3*	p.Gly1134Arg
11172.p1	60	*NR4A2*	p.Tyr275His
11660.p1	60	*NTNG1*	p.Thr135Ile
12532.p1	110	*NTNG1*	p.Tyr23Cys
11093.p1	91	*OPRL1*	p.Arg157Cys
13793.p1	56	*PCDHB4*	p.Asp555His
11707.p1	23	*PDCD1*	Frameshift indel
12304.p1	83	*PSEN1*	p.Thr421Ile
11390.p1	77	*PTEN*	p.Thr167Asn
13629.p1	63	*PTPRK*	p.Arg784His
13333.p1	69	*RGMA*	p.Val379Ile
13222.p1	86	*RPS6KA3*	p.Ser369TERM
11257.p1	128	*RUVBL1*	p.Leu365Gln
11843.p1	113	*SESN2*	p.Ala46Thr
12933.p1	41	*SETBP1*	Frameshift indel
12565.p1	79	*SETD2*	Frameshift indel
12335.p1	47	*TBL1XR1*	p.Leu282Pro
11480.p1	41	*TBR1*	Frameshift indel
11569.p1	67	*TNKS*	p.Arg568Thr
12621.p1	120	*TSC2*	p.Arg1580Trp
11291.p1	83	TSPAN17	p.Ser75TERM
11006.p1	125	*UBE3C*	p.Ser845Phe
12161.p1	95	*UBR3*	Frameshift indel
12521.p1	78	*USP15*	Frameshift indel
11526.p1	92	*ZBTB41*	p.Tyr886His
13335.p1	25	*ZNF420*	p.Leu76Pro

Proband	NVIQ	CNV Candidate gene	Type
11928.p1	66	*CHRNA7*	Duplication
13815.p1	56	*CNTNAP4*	Deletion
13726.p1	59	*CTNND1*	Deletion
12581.p1	34	*EHMT1*	Deletion
13335.p1	25	*TBX6*	Duplication

Top candidate mutations based on severity and/or supporting evidence from the literature.

interconnected network wherein 92% of gene pairs in the connected component are linked by paths of three or fewer edges (Fig. 2a). We tested this degree of interconnectivity by simulation ($n = 10{,}000$ replicates; Methods and Supplementary Fig. 9) and found that our experimental network had significantly more edges ($P < 0.0001$) and a greater clustering coefficient ($P < 0.0001$) than expected by chance.

To investigate the relevance of this network to autism further, we applied degree-aware disease gene prioritization (DADA)[23], based on the same PPI database to rank all genes based on their relatedness to a set of 103 previously identified ASD genes[17]. We found that the genes with severe mutations ranked significantly higher than all other genes (Mann–Whitney *U*-test, $P < 4.0 \times 10^{-4}$), suggesting enrichment of ASD candidates. Furthermore, the 49 members of the connected component overwhelmingly drove this difference (Mann–Whitney *U*-test, $P < 1.6 \times 10^{-8}$), as the unconnected members were not significant on their own (Mann–Whitney *U*-test, $P < 0.28$), increasing our confidence that these connected gene products are probably related to ASD (Supplementary Fig. 10). Consistent with this finding, the rankings of unaffected sibling events are highly similar to the unconnected component, strengthening our confidence in the enrichment of the connected component of proband events for ASD-relevant genes.

Members of this network have known functions in β-catenin and p53 signalling, chromatin remodelling, ubiquitination and neuronal development (Fig. 2a). A fundamental developmental regulator observed in the network is *CTNNB1* (catenin (cadherin-associated protein), β1, 88 kDa), also known as β-catenin. Interestingly, a parallel analysis using ingenuity pathway analysis (IPA) shows an enrichment of upstream interacting genes of the β-catenin pathway (8 of 358, $P = 0.0030$; see Methods, Supplementary Table 13 and Supplementary Fig. 11). A role for Wnt/β-catenin signalling in ASD was previously proposed[24], largely on the basis of the association of common variants in *EN2* and *WNT2*, and the high rate of children with macrocephaly. It is striking that both individuals with *CHD8* mutations in this study have multiple *de novo* disruptive missense mutations in this pathway or closely related pathways (Fig. 2b, c and Supplementary Fig. 12) and both have macrocephaly.

In addition, the pathway analysis shows several other disrupted genes not identified in the PPI that are involved in common pathways, which in some cases are linked to β-catenin (Supplementary Discussion and Supplementary Fig. 11). *TBR1*, for example, is a transcription factor that has a critical role in the development of the cerebral cortex[25]. TBR1 binds with CASK and regulates several candidate genes for ASD and intellectual disability including *GRIN2B*, *AUTS2* and *RELN*—genes of recurrent ASD mutation, some of which are described here and in other studies[4,9,11,17].

Our exome analysis of *de novo* coding mutations in 209 autism trios identified only two recurrently altered genes, consistent with extreme locus heterogeneity underlying ASD. This extreme heterogeneity necessitates the analysis of very large cohorts for validation. We implemented a cost-effective approach based on molecular inversion probe (MIP) technology[26] for the targeted resequencing of six candidate genes in ~2,500 individuals, including 1,703 simplex ASD probands and 744 controls. Four of these candidates (*FOXP1*, *GRIN2B*, *LAMC3* and *SCN1A*) were identified previously[4], whereas two (*FOXP2*, OMIM 602081 and *GRIN2A*, OMIM 613971) are related genes implicated in other neurodevelopmental phenotypes. We identified all previously observed *de novo* events (that is, in the same individuals), as well as additional *de novo* events in *GRIN2B* (two protein-truncating events), *SCN1A* (a missense) and *LAMC3* (a missense) (Supplementary Table 8). The observed number of *de novo* events was compared with expectations based on the mutation rates estimated for each gene (Methods and Supplementary Table 8), with *GRIN2B* showing the highest significance (uncorrected *P* value <0.0002). Notably, the three *de novo* events observed in *GRIN2B* are all predicted to be protein truncating, whereas no events truncating *GRIN2B* were found in more than 3,000 controls (Methods).

Our analysis predicts extreme locus heterogeneity underlying the genetic aetiology of autism. Under a strict sporadic disorder–*de novo* mutation model, if 20–30% of our *de novo* point mutations are considered to be pathogenic, we can estimate between 384 and 821 loci (Methods and Supplementary Fig. 13). We reach a similar estimate if we consider recurrences from ref. 9. It is clear from phenotype and genotype data that there are many 'autisms' represented under the current umbrella of ASD and other genetic models are more likely in different contexts (for example, families with multiple affected

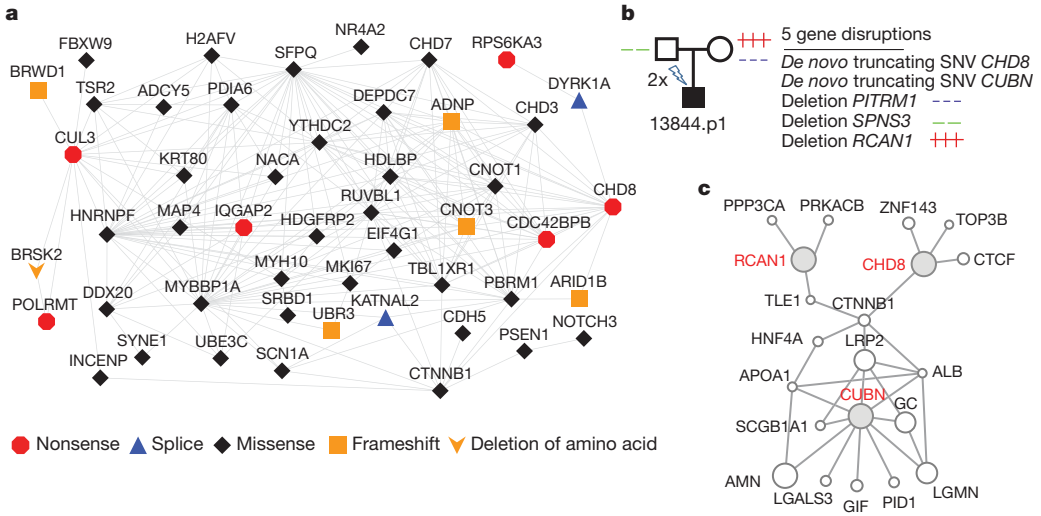


Figure 2 | Mutations identified in protein–protein interaction (PPI) networks. **a**, The 49-gene connected component of the PPI network formed from 126 genes with severe *de novo* mutations among the 209 probands. **b**, Proband 13844 inherits three rare gene-disruptive CNVs and carries two *de novo* truncating mutations. **c**, GeneMANIA[22] view of three of the affected genes (**b**) (red labels) which encode proteins that are part of a β-catenin-linked network. This proband is macrocephalic, impaired cognitively, and has deficits in social behaviour and language development (Supplementary Discussion).

individuals). There is marked convergence on genes previously implicated in intellectual disability and developmental delay. As has been noted for CNVs, this indicates that nosological divisions may not readily translate into differences at the molecular level. We believe that there is value in comparing mutation patterns in children with developmental delay (without features of autism) to those in children with ASD.

Although there is no one major genetic lesion responsible for ASD, it is still largely unknown whether there are subsets of individuals with a common or strongly related molecular aetiology and how large these subsets are likely to be. Using gene expression, protein–protein interactions, and CNV pathway analysis, recent reports have highlighted the role of synapse formation and maintenance[27–29]. We find it intriguing that 49 proteins found to be mutated here have critical roles in fundamental developmental pathways, including β-catenin and p53 signalling, and that patients have been identified with multiple disruptive *de novo* mutations in interconnected pathways. The latter observations are consistent with an oligogenic model of autism where both *de novo* and extremely rare inherited SNV and CNV mutations contribute in conjunction to the overall genetic risk. Recent work has supported a role for these interconnected pathways in neuronal stem-cell fate-determination, differentiation and synaptic formation in humans and animal models[24,30,31]. Given that fundamental developmental processes have previously been found to underlie syndromic forms of autism, a wider role of these pathways in idiopathic ASD would not be entirely surprising and would help explain the extreme genetic heterogeneity observed in this study.

METHODS SUMMARY

Exome capture, alignments and base-calling. Genomic DNA was derived directly from whole blood. Exomes were considered to be completed when ~90% of the capture target exceeded 8-fold coverage and ~80% exceeded 20-fold coverage. Exomes for the 189 trios (and 31 unaffected siblings) were captured with NimbleGen EZ Exome V2.0. Reads were mapped as in ref. 4 to a custom reference genome assembly (GRC build37). Genotypes were generated with GATK unified genotyper and parallel SAMtools pipeline[4]. Exomes for the unaffected siblings matching the pilot trios were captured and analysed as in ref. 4. Predicted *de novo* events were called as in ref. 4 and confirmed by capillary sequencing in all family members (for 176 of the 189 trios, this also included one unaffected sibling). Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid.

Exome read-depth CNV analysis. Reads were mapped using mrsFAST and normalized reads per kilobase of exon per million mapped reads (RPKM) values calculated by exon. Population normalization was performed using a set of 366 non-ASD exomes. Calls were made if three or more exons passed a threshold value and cross-validated calls using two orthogonal platforms, custom array CGH and Illumina 1M array data[2]. CNVs were filtered to identify *de novo* and rare inherited events by comparison with 2,090 controls and 1,651 parent profiles.

Network reconstruction and null model estimation. PPI networks were generated using physical interaction data from GeneMANIA[22]. Null models were estimated using gene-specific mutation rate estimates based on human–chimp divergence. To rank candidate genes we obtained the seed ASD list from ref. 17 and severe disruptive *de novo* events from all families ($n = 209$). Given the PPI network and seed gene product list, we used DADA[23] for ranking each gene.

Human subjects. All samples and phenotypic data were collected under the direction of the Simons Simplex Collection by its 12 research clinic sites (http://sfari.org/sfari-initiatives/simons-simplex-collection). Parents consented and children assented as required by each local institutional review board. Participants were de-identified before distribution. Research was approved by the University of Washington Human Subject Division under non-identifiable biological specimens/data.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 8 September 2011; accepted 23 February 2012.
Published online 4 April 2012.

1. Schaaf, C. P. & Zoghbi, H. Y. Solving the autism puzzle a few pieces at a time. *Neuron* **70,** 806–808 (2011).
2. Sanders, S. J. *et al.* Multiple recurrent *de novo* CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70,** 863–885 (2011).
3. Levy, D. *et al.* Rare *de novo* and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886–897 (2011).
4. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature Genet.* **43,** 585–589 (2011).
5. Hultman, C. M., Sandin, S., Levine, S. Z., Lichtenstein, P. & Reichenberg, A. Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies. *Mol. Psychiatry* **16,** 1203–1212 (2010).
6. Fischbach, G. D. & Lord, C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68,** 192–195 (2010).
7. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107,** 961–968 (2010).
8. Xu, B. *et al.* Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nature Genet.* **43,** 864–868 (2011).
9. Sanders, S. J. *et al. De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* http://dx.doi.org/10.1038/nature10945 (this issue).
10. Hehir-Kwa, J. Y. *et al. De novo* copy number variants associated with intellectual disability have a paternal origin and age bias. *J. Med. Genet.* **48,** 776–778 (2011).
11. O'Roak, B. J. & State, M. W. Autism genetics: strategies, challenges, and opportunities. *Autism Res.* **1,** 4–17 (2008).

12. Nishimura-Akiyoshi, S., Niimi, K., Nakashiba, T. & Itohara, S. Axonal netrin-Gs transneuronally determine lamina-specific subdendritic segments. *Proc. Natl Acad. Sci. USA* **104,** 14801–14806 (2007).
13. Borg, I. *et al.* Disruption of Netrin G1 by a balanced chromosome translocation in a girl with Rett syndrome. *Eur. J. Hum. Genet.* **13,** 921–927 (2005).
14. Nishiyama, M. *et al.* CHD8 suppresses p53-mediated apoptosis through histone H1 recruitment during early embryogenesis. *Nature Cell Biol.* **11,** 172–182 (2009).
15. Thompson, B. A., Tremblay, V., Lin, G. & Bochar, D. A. CHD8 is an ATP-dependent chromatin remodeling factor that regulates β-catenin target genes. *Mol. Cell. Biol.* **28,** 3894–3904 (2008).
16. Batsukh, T. *et al.* CHD8 interacts with CHD7, a protein which is mutated in CHARGE syndrome. *Hum. Mol. Genet.* **19,** 2858–2866 (2010).
17. Betancur, C. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain Res.* **1380,** 42–77 (2011).
18. Moller, R. S. *et al.* Truncation of the Down syndrome candidate gene *DYRK1A* in two unrelated patients with microcephaly. *Am. J. Hum. Genet.* **82,** 1165–1170 (2008).
19. Cooper, G. M. *et al.* A copy number variation morbidity map of developmental delay. *Nature Genet.* **43,** 838–846 (2011).
20. Buysse, K. *et al.* Delineation of a critical region on chromosome 18 for the del(18)(q12.2q21.1) syndrome. *Am. J. Med. Genet. A.* **146A,** 1330–1334 (2008).
21. Hoischen, A. *et al. De novo* mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genet.* **42,** 483–485 (2010).
22. Warde-Farley, D. *et al.* The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38,** W214–W220 (2010).
23. Erten, S., Bebek, G., Ewing, R. & Koyutürk, M. DADA: Degree-aware algorithms for network-based disease gene prioritization. *BioData Mining* **4,** 19 (2011).
24. De Ferrari, G. V. & Moon, R. T. The ups and downs of Wnt signaling in prevalent neurological disorders. *Oncogene* **25,** 7545–7553 (2006).
25. Bedogni, F. *et al.* Tbr1 regulates regional and laminar identity of postmitotic neurons in developing neocortex. *Proc. Natl Acad. Sci. USA* **107,** 13129–13134 (2010).
26. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6,** 315–316 (2009).
27. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474,** 380–384 (2011).
28. Sakai, Y. *et al.* Protein interactome reveals converging molecular pathways among autism disorders. *Sci. Transl. Med.* **3,** 86ra49 (2011).
29. Gilman, S. R. *et al.* Rare *de novo* variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70,** 898–907 (2011).
30. Ille, F. & Sommer, L. Wnt signaling: multiple functions in neural development. *Cell. Mol. Life Sci.* **62,** 1100–1108 (2005).
31. Tedeschi, A. & Di Giovanni, S. The non-apoptotic role of p53 in neuronal biology: enlightening the dark side of the moon. *EMBO Rep.* **10,** 576–583 (2009).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank and recognize the following ongoing studies that produced and provided exome variant calls for comparison: NHLBI Lung Cohort Sequencing Project (HL 1029230), NHLBI WHI Sequencing Project (HL 102924), NIEHS SNPs (HHSN273200800010C), NHLBI/NHGRI SeattleSeq (HL 094976), and the Northwest Genomics Center (HL 102926). We are grateful to all of the families at the participating Simons Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, E. Hanson, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, K. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, Z. Warren and E. Wijsman). We also acknowledge M. State and the Simons Simplex Collection Genetics Consortium for providing Illumina genotyping data, T. Lehner and the Autism Sequencing Consortium for providing an opportunity for pre-publication data exchange among the participating groups. We appreciate obtaining access to phenotypic data on SFARI Base. This work was supported by the Simons Foundation Autism Research Initiative (SFARI 137578 and 191889; E.E.E., J.S. and R.B.) and NIH HD065285 (E.E.E. and J.S.). E.B. is an Alfred P. Sloan Research Fellow. E.E.E. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions E.E.E., J.S. and B.J.O. designed the study and drafted the manuscript. E.E.E. and J.S. supervised the study. R.B., B.R. and B.J.O. analysed the clinical information. R.B., L.V., S.G., E.K., N.K. and B.P.C. contributed to the manuscript. S.G., N.K., B.P.C., A.K., C.B., M.M. and L.V. generated and analysed CNV data. B.J.O. and L.V. performed MIP resequencing and mutation validations. I.B.S., E.H.T., B.J.O. and J.S. developed MIP protocol and analysis. B.V. and J.M.A. generated loci-specific mutation rate estimates. R.L. and E.B. performed PPI network analysis and simulations. E.K. performed DADA analysis. C.L. performed Illumina sequencing. J.D.S., I.B.S., E.H.T. and C.L. analysed sequence data. B.P.C. performed IPA analysis. B.J.O., E.K. and N.K. developed the *de novo* analysis pipelines and analysed sequence data. D.A.N., M.J.R., J.D.S. and E.H.T. supervised exome sequencing and primary analysis.

Author Information Access to the raw sequence reads can be found at the NCBI database of Genotypes and Phenotypes (dbGaP) and National Database for Autism Research under accession numbers phs000482.v1.p1 and NDARCOL0001878, respectively. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.E.E. (eee@gs.washington.edu) or J.S. (shendure@uw.edu).

METHODS

Exome capture, alignments and base-calling. Exomes for the 189 trios (and 31 unaffected siblings) were captured with NimbleGen EZ Exome V2.0. Final libraries were then sequenced on either an Illumina GAIIx (paired- or single-end 76-bp reads) or HiSeq2000 (paired- or single-end 50-bp reads). Reads were mapped to a custom GRCh37/hg19 build using BWA 0.5.6 (ref. 32). Read qualities were recalibrated using GATK Table Recalibration 1.0.2905 (ref. 33). Picard-tools 1.14 was used to flag duplicate reads (http://picard.sourceforge.net/). GATK IndelRealigner 1.0.2905 was used to realign reads around insertion/deletion (indel) sites. Genotypes were generated with GATK Unified Genotyper[33] with FILTER = "QUAL ≤ 50.0 || AB ≥ 0.75 || HRun > 3 || QD < 5.0" and in parallel with the SAMtools pipeline as described previously[4]. Only positions with at least eightfold coverage were considered. All pilot sibling exomes were captured and analysed as described previously[4]. Predicted *de novo* events were called and compared against a set of 946 other exomes to remove recurrent artefacts and likely undercalled sites. Indels were also called with the GATK Unified Genotyper and SAMtools and filtered to those with at least 25% of reads showing a variant at a minimum depth of 8×. Mutations were phased using molecular cloning of PCR fragments, read-pair information, linked informative SNPs, and obligate carrier status. To identify rare private variants (singleton), the full variant list was compared against a larger set of 1,779 other exomes. Predicted *de novo* indels were also filtered against this larger set.

Sanger validations. All reported *de novo* events (exome or MIP capture) were validated by designing primers with BatchPrimer3 followed by PCR amplification and Sanger sequencing. We performed PCR reactions using 10 ng of DNA from father, mother, unaffected sibling (when available), and proband and performed Sanger capillary sequencing of the PCR product using forward and reverse primers. In some cases, one direction could not be assessed due to the presence of repeat elements or indels in close proximity to the mutation event.

Mutation candidate gene analysis. We examined whether each non-synonymous or CNV *de novo* event may be contributing to the aetiology of ASD by evaluating the likelihood deleteriousness of the change (GERP, Grantham score) and intersecting with known syndromic and non-syndromic candidate genes, CNV morbidity maps, and information in OMIM and PubMed. Mutations were considered severe if they were truncating, missense with Grantham score ≥50 and GERP score ≥3 or only Grantham score ≥85, or deleted a highly conserved amino acid. For genes that had not previously been implicated in ASD, we gave priority to those with structural similarities to known candidate or strong evidence of neural function or development.

Exome read-depth CNV discovery. To find CNVs using exome read-depth data, we first mapped sequenced reads to the hg19 exome using the mrsFAST aligner[34]. Next, we applied a novel method (N.K. *et al.*, manuscript in preparation), which uses normalized RPKM values[35] of the ~194,000 captured exons/sequences, subsequent population normalization using 366 exomes from the Exome Sequencing Project and singular value decomposition to remove systematic bias present within exome capture reactions. Rare CNVs were detected using a threshold cutoff of the normalized RPKM values, and we required at least three exons above our threshold in order to make a call. We made a total of 1,077 deletion or duplication calls in 366 individuals (range 0–14, median = 3, mean = 2.94).

CNV detection using array CGH. A custom-targeted 2 × 400K Agilent chip with median probe spacing of 500 bp in the genomic hotspots flanked by segmental duplications or Alu repeats and probe spacing of 14 kb in the genomic backbone was designed. All experiments were performed according to the manufacturer's instructions using NA12878 as the female reference and NA18507 as the male reference (Coriell). Data analysis was performed following feature extraction using DNA analytics with ADM-2 setting. All CNV calls were visually inspected in the UCSC Genome Browser. CNV calls from probands were then intersected with those from parents and also with 377 controls recruited through NIMH Genetics Initiative[36,37] and ClinSeq cohort[38] analysed on the same microarray platform. The NIMH set of controls were ascertained by the NIMH Genetics Initiative[36] through an online self-report based on the Composite International Diagnostic Instrument Short-Form (CIDI-SF)[37]. Those who did not meet DSM-IV criteria for major depression, denied a history of bipolar disorder or psychosis, and reported exclusively European origins were included[39,40]. Samples from the ClinSeq cohort were selected from a population representing a spectrum of atherosclerotic heart disease[38]. *De novo* and inherited potential pathogenic CNVs were selected only if they intersected with RefSeq coding sequence and allowing for a frequency of <1% in the controls and <50% segmental duplication content.

Illumina array CNV calling. CNV calling was performed in hg18 as described previously[41], using an HMM that incorporates both allele frequencies (BAF) and total intensity values (logR). In total, we generated CNV calls for 841 probands, 1,651 parents and 793 siblings including the samples reported recently[2]. Of the 122 families selected for CNV comparisons in this study, calls were generated for 107 probands. Of these, both parents were profiled for 101 families and one parent was profiled for the remaining six families. In addition, at least one sibling was profiled for 99 of these families.

Independent of array CGH detection, to identify putatively pathogenic CNVs, we first compared our data to 2,090 control samples derived from the Wellcome Trust Case Control Consortium (WTCCC) National Blood Services Cohort[19,42] and filtered all CNVs present in 1% (20) of WTCCC2 controls or 1% (16) of parents by 50% reciprocal overlap with matching copy number status. In addition, similar to the filtering criteria used for array CGH detection, we selected only CNVs that contained less than 50% segmental duplication and intersected with RefSeq coding sequence. To select putative *de novo* CNVs, we further required the CNV not to be present in family-matched parents and siblings. Additionally, we filtered CNVs present in >0.1% (2) of the full 1,651 parent set. To select potential, rare inherited events, we required the CNV be detected in a matched parent or sibling. Finally, we filtered the genes inside each CNV under the same criteria (to account for smaller or larger CNPs) and removed CNVs with no remaining genes.

CNV cross validation. High-confidence, cross-validated *de novo* and inherited CNVs were selected by identifying events detected by at least two of three methodologies. To account for the variable breakpoint definitions in array CGH, SNP arrays, and exome copy number profiles, we aligned the CNVs by at least one overlapping gene ID and reported each CNV region by its maximal outer boundaries. This identified six *de novo* and 70 rare inherited events for further study (Supplementary Table 6).

Ingenuity pathway analysis. Ingenuity pathway analysis (IPA) was performed to identify potential functional enrichments within both our PPI (49 genes) and overall set of 126 genes. RefSeq reference gene list was used as a background list for all analysis. To confirm our results pertaining to *CTNNB1* upstream enrichment, we simulated 10,000 random populations of 209 individuals using Poisson priors for each gene based on their estimated mutation rates (see below), with a global correction factor resulting in selecting a mean of 126 genes per population. We then used this simulation data to calculate the probability of observing eight direct upstream interactors of *CTNNB1* and determined that our data set is enriched for these genes with $P = 0.0030$.

Estimating locus-specific mutation rates. Human–chimpanzee alignments were downloaded from the UCSC Genome Browser (reference versions GRCb37 and panTro2, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/syntenicNet/). The more conservative syntenicNet alignments were used (details in http://hgdownload.cse.ucsc.edu/goldenPath/hg19/vsPanTro2/README.txt). Gene definitions were downloaded from the UCSC Table Browser, from the RefSeq Genes track, and the refFlat table. Exons were extended by 2 bp, and overlapping exons were merged using BEDTools. Non-exonic sequence was not considered. For each gene, we extracted: (1) d = the number of differences between chimpanzee and human; and (2) n = the number of bases aligned. We assumed a divergence time between human and chimpanzee of 12 million years (Myr) and an average generation time of 25 years. We then calculated gene-specific mutation rates per site per generation: $r = (d/n)/(12\ \text{Myr}/25\ \text{years/generation})$. We calculated the probability of observing X+ events using the Poisson distribution defined by the number of chromosomes screened and the size of the coding region, including actual splice bases.

Network simulation and null model estimation. To generate a null distribution of gene mutations, *de novo* mutation rates were estimated from human–chimp mutation rates. A pseudocount of 2.0833×10^{-6} (the smallest calculated in the gene set) was applied to any exon with a mutation rate of zero. To create null gene sets, genes were drawn uniformly from this background distribution. Human protein–protein interaction data were collected from GeneMANIA[22] on 29 August 2011. Only direct physical interactions from the *Homo sapiens* database were considered. The list comprises approximately 1.5 million physical interactions, gathered from 150 studies. A protein interaction network was created from each experimental and null gene set by drawing edges between genes with physical interactions reported in the GeneMANIA database. Qualitatively similar results were achieved by including only interactions supported by multiple independent data sources. For each network, clustering coefficient, centralization, average shortest path length, density, and heterogeneity were determined using Cytoscape[43] and Network Analyzer[44]. Duplicate- and self-interactions were not considered in calculating network statistics.

Disease gene prioritization based on PPI networks. We applied degree-aware algorithms to rank a set of candidate genes with respect to a set of products of genes associated with ASD using human PPI networks. We used the integrated human PPI network data collected from GeneMANIA[22] on 29 August 2011. The PPI network contains 12,007 proteins with ~1.5 million direct physical interactions associated with a reliability score. We obtain the seed proteins for the ASD from the list of ref. 17. For the candidate set we used 126 gene products from the severe

disruptive *de novo* events from the pilot autism project[4] and the current study. Given the GeneMANIA PPI network and Betancur seed gene product list, we used DADA[23] for ranking the candidate genes. We emphasize that this ranking is not implying causality but rather relatedness to genes previously and independently associated with ASD. For testing the significance of this ranking, we rank all the gene products except the seed set using the same algorithm. On the basis of the ranking result, we applied a Mann–Whitney *U* rank sum test (one-tailed) on the candidate set compared to all the other genes.

MIP protocol. Each of 1,703 autism probands from the SSC collection and 744 controls from the NIMH collection was subjected to MIP-based multiplex capture of the six genes: *SCN1A*, *GRIN2B*, *GRIN2A*, *LAMC3*, *FOXP1* and *FOXP2*. For each library, 50 ng of DNA was used. Individually synthesized 70 mer MIPs ($n = 355$) were pooled and 5′ phosphorylated with T4 PNK (NEB). Hybridization with MIPs, gap filling and ligation were performed in one step for 45–48 h at 60 °C, followed by an exonuclease treatment of 30 min at 37 °C, similar to ref. 45, with modifications for reduced MIP number (B.J.O. *et al.*, manuscript in preparation). Amplification of the library was performed by PCR using different barcoded primers for each library. Then barcoded libraries were pooled, purified using Agencourt AMPure XP and one lane of 101-bp paired-end reads was generated for each mega-pool (~384) on an Illumina HiSeq 2000 according to manufacturer's instructions. Raw reads were mapped to the genome as in ref. 4. MIP targeting arms were then removed and variants called using SAMtools[4]. A 25-fold coverage, with AB allele ration <0.7, and quality 30 threshold was used for high-confident variant calling. Private (possible *de novo*) variants were identified by filtering against 1,779 other exomes. The parents of children with disruptive rare variants were then captured. Variants not seen or with low coverage in the parents were validated by Sanger capillary-based fluorescent sequencing. No truncating variants of *GRIN2B* were observed in the MIP sequenced controls or the Exome Variant Server ESP2500 release (NHLBI Exome Sequencing Project (ESP), Seattle, Washington, http://evs.gs.washington.edu/EVS/).

Estimating the number of autism loci. The gene-level specificity of exome sequencing enables the estimation of the number of recurrently mutated genes implicated in the genetic aetiology of sporadic ASD. This question can be reformulated as the 'unseen species problem' (see ref. 46 for review and ref. 2 for application to *de novo* CNVs discovered in autism), where genes with severe *de novo* events in probands are considered 'observed species', and binned by their frequency of appearance (that is, singletons, doubletons, etc.). We estimated the total number of genes implicated in autism (the total number of species) using several different estimators (implemented in the R package SPECIES, http://www.jstatsoft.org/), as well as the formula provided in ref. 2. This estimate depends on the number of singletons and twin pairs of genes observed in probands, as well as the fraction of *de novo* events believed to be pathogenic for autism, that is, single, disruptive events that can cause autism on their own. We assumed that both of our recurrent severe *de novo* events (affecting *CHD8* and *NTNG1*) were pathogenic; these compose the entire set of twin pairs. The number of singletons is based on the estimated a priori fraction of the observed events that are pathogenic for autism. Across this sliding scale, the estimated number of loci is plotted in Supplementary Fig. 13. For example, using the estimator from ref. 47, if 20–50% of our *de novo* severe events are considered pathogenic, exome sequencing of a large number of additional samples would reveal between 182 and 992 pathogenic genes harbouring coding *de novo* point mutations (Supplementary Fig. 13); if all the observed severe *de novo* events in our experiment are included as pathogenic singletons, the number of implicated loci increases to more than 3,000.

32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
33. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43** (2011).
34. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nature Methods* **7,** 576–577 (2010).
35. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
36. Moldin, S. O. NIMH Human Genetics Initiative: 2003 update. *Am. J. Psychiatry* **160,** 621–622 (2003).
37. Kessler, R. C. & Ustun, T. B. The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int. J. Methods Psychiatr. Res.* **13,** 93–121 (2004).
38. Biesecker, L. G. *et al.* The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res.* **19,** 1665–1674 (2009).
39. Talati, A., Fyer, A. J. & Weissman, M. M. A comparison between screened NIMH and clinically interviewed control samples on neuroticism and extraversion. *Mol. Psychiatry* **13,** 122–130 (2008).
40. Baum, A. E. *et al.* A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol. Psychiatry* **13,** 197–207 (2008).
41. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84,** 148–161 (2009).
42. Craddock, N. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464,** 713–720 (2010).
43. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27,** 431–432 (2011).
44. Assenov, Y., Ramirez, F., Schelhorn, S. E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24,** 282–284 (2008).
45. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7,** 111–118 (2010).
46. Bunge, J. & Fitzpatrick, M. Estimating the number of species - a Review. *J. Am. Stat. Assoc.* **88,** 364–373 (1993).
47. Chao, A. & Lee, S. M. Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **87,** 210–217 (1992).

LETTER

doi:10.1038/nature10992

Mitochondrial DNA that escapes from autophagy causes inflammation and heart failure

Takafumi Oka[1], Shungo Hikoso[1], Osamu Yamaguchi[1], Manabu Taneike[1,2], Toshihiro Takeda[1], Takahito Tamai[1], Jota Oyabu[1], Tomokazu Murakawa[1], Hiroyuki Nakayama[3], Kazuhiko Nishida[1,2], Shizuo Akira[4,5], Akitsugu Yamamoto[6], Issei Komuro[1] & Kinya Otsu[1,2]

Heart failure is a leading cause of morbidity and mortality in industrialized countries. Although infection with microorganisms is not involved in the development of heart failure in most cases, inflammation has been implicated in the pathogenesis of heart failure[1]. However, the mechanisms responsible for initiating and integrating inflammatory responses within the heart remain poorly defined. Mitochondria are evolutionary endosymbionts derived from bacteria and contain DNA similar to bacterial DNA[2–4]. Mitochondria damaged by external haemodynamic stress are degraded by the autophagy/lysosome system in cardiomyocytes[5]. Here we show that mitochondrial DNA that escapes from autophagy cell-autonomously leads to Toll-like receptor (TLR) 9-mediated inflammatory responses in cardiomyocytes and is capable of inducing myocarditis and dilated cardiomyopathy. Cardiac-specific deletion of lysosomal deoxyribonuclease (DNase) II showed no cardiac phenotypes under baseline conditions, but increased mortality and caused severe myocarditis and dilated cardiomyopathy 10 days after treatment with pressure overload. Early in the pathogenesis, DNase II-deficient hearts showed infiltration of inflammatory cells and increased messenger RNA expression of inflammatory cytokines, with accumulation of mitochondrial DNA deposits in autolysosomes in the myocardium. Administration of inhibitory oligodeoxynucleotides against TLR9, which is known to be activated by bacterial DNA[6], or ablation of *Tlr9* attenuated the development of cardiomyopathy in DNase II-deficient mice. Furthermore, *Tlr9* ablation improved pressure overload-induced cardiac dysfunction and inflammation even in mice with wild-type *Dnase2a* alleles. These data provide new perspectives on the mechanism of genesis of chronic inflammation in failing hearts.

Mitochondrial DNA has similarities to bacterial DNA, which contains inflammatogenic unmethylated CpG motifs[2–4,7,8]. Damaged mitochondria are degraded by autophagy, which involves the sequestration of cytoplasmic contents in a double-membraned vacuole, the autophagosome and the fusion of the autophagosome with the lysosome[9]. Pressure overload induces the impairment of mitochondrial cristae morphology and functions in the heart[10,11]. We have previously reported that autophagy is an adaptive mechanism to protect the heart from haemodynamic stress[5].

DNase II, encoded by *Dnase2a*, is an acid DNase found in the lysosome[12]. DNase II in macrophages has an essential role in the degradation of the DNA of apoptotic cells after macrophages engulf them[13]. In the present study, we hypothesized that DNase II in cardiomyocytes digests mitochondrial DNA in the autophagy system to protect the heart from inflammation in response to haemodynamic stress.

First, we examined the alteration of DNase II activity in the heart in response to pressure overload. In wild-type mice, pressure overload by thoracic transverse aortic constriction (TAC) induced cardiac hypertrophy 1 week after TAC and heart failure 8–10 weeks after TAC[5]. DNase II activity was upregulated in hypertrophied hearts, but not in failing hearts (Supplementary Fig. 1a). Immunohistochemical analysis showed infiltration of $CD45^+$ leukocytes, including $CD68^+$ macrophages in failing hearts (Supplementary Fig. 1b). Then, we stained the heart sections with PicoGreen[14], anti-LAMP2a and anti-LC3 (ref. 15) antibodies, which was used for the detection of DNA, lysosomes and autophagosomes, respectively (Supplementary Figs 1c, d and 2a). We observed PicoGreen- and LAMP2a-positive deposits and PicoGreen- and LC3-positive deposits in failing hearts, but not in hypertrophied hearts, suggesting the accumulation of DNA in autolysosomes in failing hearts.

We crossed mice bearing a $Dnase2a^{flox}$ allele[13] with transgenic mice expressing *Cre* recombinase under the control of the α-myosin heavy chain promoter (α-MyHC)[16], to produce $Dnase2a^{flox/flox}$;α-MyHC-Cre^+ ($Dnase2a^{-/-}$) mice. We used $Dnase2a^{flox/flox}$;α-MyHC-Cre^- littermates ($Dnase2a^{+/+}$) as controls. The resulting $Dnase2a^{-/-}$ mice were born at the expected Mendelian frequency. In $Dnase2a^{-/-}$ mice, we observed a 90.1% reduction in the level of *Dnase2a* messenger RNA (mRNA) and a 95.1% decrease in DNase II activity in purified adult cardiomyocyte preparation (Supplementary Fig. 3a, b). Physiological parameters and basal cardiac function assessed by echocardiography showed no differences between $Dnase2a^{+/+}$ and $Dnase2a^{-/-}$ mice (Supplementary Table 1). These results indicate that DNase II does not appear to be required during normal embryonic development or for normal heart growth in the postnatal period.

To clarify the role of DNase II in cardiac remodelling, $Dnase2a^{-/-}$ mice were subjected to TAC. DNase II activity was upregulated in response to pressure overload in $Dnase2a^{+/+}$ hearts and was lower in sham- and TAC-operated $Dnase2a^{-/-}$ hearts than that in the corresponding controls (Supplementary Fig. 3c). Twenty-eight days after TAC, 57.1% of $Dnase2a^{-/-}$ mice had died, whereas 85.7% of $Dnase2a^{+/+}$ mice were still alive (Fig. 1a). The $Dnase2a^{-/-}$ hearts showed left ventricular dilatation and severe contractile dysfunction 10 days after TAC (Fig. 1b–d and Supplementary Table 2). The lung-to-body weight ratio, an index of lung congestion, was elevated in TAC-operated $Dnase2a^{-/-}$ mice (Fig. 1e). The increases in the heart-to-body weight ratio and cardiomyocyte cross-sectional area by TAC were larger in $Dnase2a^{-/-}$ mice than in $Dnase2a^{+/+}$ mice (Fig. 1e, f). TAC-operated $Dnase2a^{-/-}$ hearts showed massive cell infiltration (Fig. 1f). Immunohistochemical analysis of the hearts showed infiltration of $CD45^+$ leukocytes, including $CD68^+$ macrophages (Supplementary Fig. 4a). The mRNA level of interleukin (IL)-6 (*Il6*) was upregulated, but not other cytokine mRNAs in TAC-operated $Dnase2a^{-/-}$ hearts (Supplementary Fig. 4b). TAC-operated $Dnase2a^{-/-}$ hearts showed intermuscular and perivascular fibrosis with increased

[1]Department of Cardiovascular Medicine, Osaka University Graduate School of Medicine, Suita, Osaka 565-0871, Japan. [2]Cardiovascular Division, King's College London, London SE5 9NU, UK. [3]Department of Clinical Pharmacology and Pharmacogenomics, Graduate School of Pharmaceutical Sciences, Osaka University, Suita, Osaka 565-0871, Japan. [4]Laboratory of Host Defense, WPI Immunology Frontier Research Center, Osaka University, Suita, Osaka 565-0871, Japan. [5]Department of Host Defense, Research Institute for Microbial Diseases, Osaka University, Suita, Osaka 565-0871, Japan. [6]Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan.

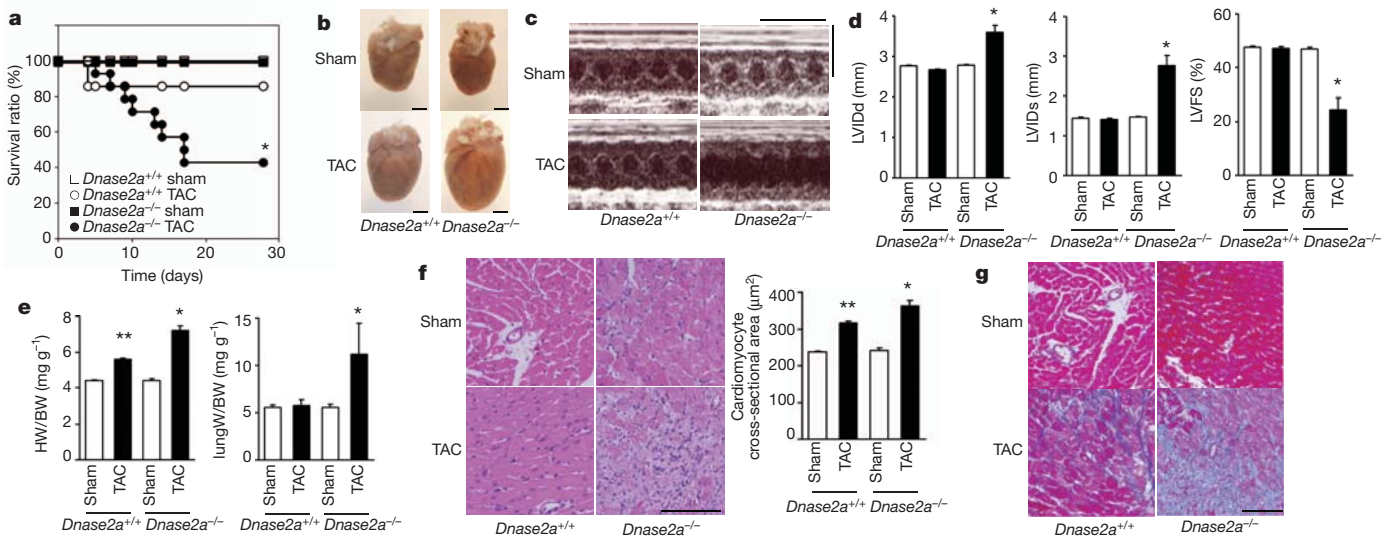


Figure 1 | TAC-induced cardiomyopathy in *Dnase2a*$^{-/-}$ mice. **a**, Survival ratio after TAC ($n = 7$–14 per group). **b**–**g**, Ten days after TAC. **b**, Gross appearance of hearts. Scale bar, 2 mm. **c**, Echocardiography. Scale bars, 0.2 s and 5 mm. Echocardiographic (**d**) and physiological (**e**) parameters ($n = 7$–13 per group). LVIDd and LVIDs, end-diastolic and end-systolic left ventricular internal dimension, respectively; LVFS, left venticular fractional shortening; HW/BW, heart/body weight. Haematoxylin and eosin-stained (**f**) and azocarmine and aniline blue (AZAN)-Mallory-stained (**g**) heart sections. Scale bar, 100 μm. Data are mean ± s.e.m. $*P < 0.05$ versus all other groups, $**P < 0.05$ versus sham-operated controls.

mRNA expression of α2 type I collagen (*Col1a2*) (Fig. 1g and Supplementary Fig. 3d). Ultrastructural analysis of TAC-operated *Dnase2a*$^{-/-}$ hearts showed a disorganized sarcomere structure, misalignment and aggregation of mitochondria, and aberrant electron-dense structures (Supplementary Fig. 4c). The mRNA levels of atrial natriuretic factor (*Nppa*) and brain natriuretic peptide (*Nppb*) were higher in TAC-operated *Dnase2a*$^{-/-}$ mice than in TAC-operated *Dnase2a*$^{+/+}$ mice (Supplementary Fig. 3d). These data suggest that DNase II plays an important role in preventing pressure overload-induced heart failure and myocarditis.

To clarify the molecular mechanisms underlying the cardiac abnormalities observed in *Dnase2a*$^{-/-}$ mice, we evaluated the phenotypes in the earlier time course after pressure overload. Chamber dilation and cardiac dysfunction developed with time after TAC in *Dnase2a*$^{-/-}$ mice (Supplementary Fig. 5a). We chose to perform the analysis 2 days after TAC to minimize the contributions of operation-related events and phenomena secondary to the initial and essential molecular event that induced cardiomyopathy. TAC-operated *Dnase2a*$^{-/-}$ hearts showed cell infiltration without apparent fibrosis (Fig. 2a, b) and infiltration of CD68^{+} macrophages and Ly6G^{+} cells (Fig. 2c). We detected increases in the mRNA levels of IL-1β (*Il1b*) and *Il6*, but not interferon-β (*Ifnb1*) and -γ (*Ifng*) or tumour-necrosis factor (TNF)-α in TAC-operated *Dnase2a*$^{-/-}$ hearts (Supplementary Fig. 6a). To identify the source of IL-1β and IL-6, we performed *in situ* hybridization analysis in heart sections. *Il1b* and *Il6* mRNA-positive cardiomyocytes were evident in TAC-operated *Dnase2a*$^{-/-}$ hearts (Supplementary Fig. 4d).

Ultrastructural analysis showed aberrant electron-dense deposits without apparent changes in sarcomeric and mitochondrial structures

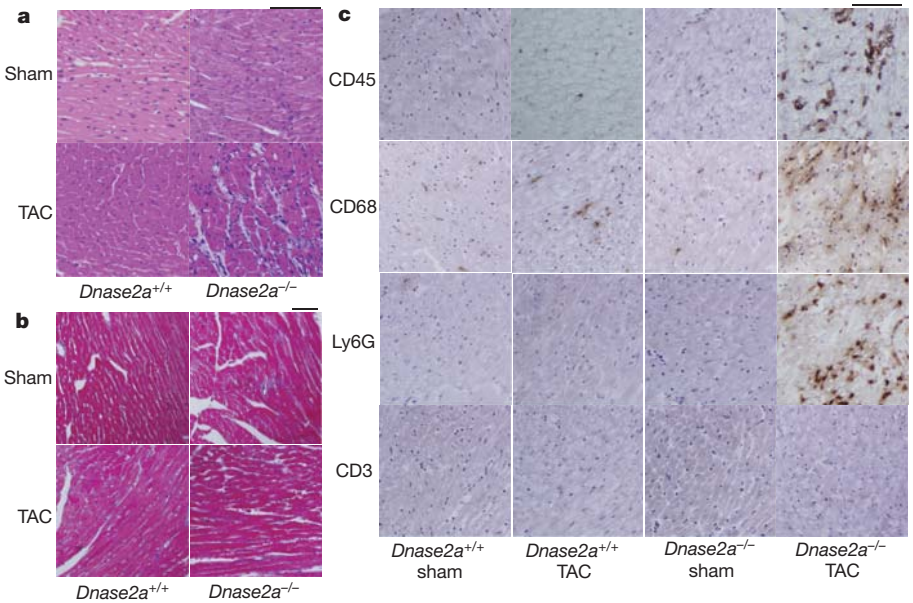


Figure 2 | Pressure overload-induced inflammatory responses in *Dnase2a*$^{-/-}$ mice 2 days after TAC. Mice were analysed 2 days after TAC (**a**–**c**). **a**, Haematoxylin and eosin-stained heart sections. Scale bar, 100 μm. **b**, AZAN-Mallory-stained sections. Scale bar, 100 μm. **c**, Immunohistochemical analysis using antibodies to CD45, CD68, Ly6G and CD3. Scale bar, 100 μm.

in TAC-operated *Dnase2a*$^{-/-}$ hearts (Fig. 3a). At higher magnification, the electron-dense deposits appeared to be autolysosomes (Fig. 3a). Immunoelectron microscopic analysis using anti-DNA antibody showed DNA deposition in autolysosomes (Fig. 3b). In TAC-operated *Dnase2a*$^{-/-}$ hearts, we observed PicoGreen- and LAMP2a-positive deposits and PicoGreen- and LC3-positive deposits (Supplementary Figs 2b and 6b, c). The PicoGreen-positive deposits were not TdT-mediated dUTP nick end labelling (TUNEL)-positive (Supplementary Fig. 6d), indicating that the DNA was not derived from fragmented nuclear DNA. To label mitochondrial DNA, mice were injected with 5-ethynyl-2′-deoxyuridine (EdU) five times before TAC. EdU, a nucleoside analogue to thymidine, is incorporated into DNA during active DNA synthesis[17]. EdU specifically binds to mitochondrial DNA during its active DNA synthesis in non-dividing cardiomyocytes. In TAC-operated *Dnase2a*$^{-/-}$ hearts, we observed EdU- and LAMP2a-positive deposits and EdU- and LC3-positive deposits (Fig. 3c, d and Supplementary Fig. 2c), indicating that mitochondrial DNA accumulated in autolysosomes.

The innate immune system is the major contributor to acute inflammation induced by microbial infection[18]. TLR9, localized in the endolysosome, senses DNA with unmethylated CpG motifs derived from bacteria and viruses. Mitochondrial DNA activates polymorphonuclear neutrophils through CpG/TLR9 interactions[19]. Immunohistochemical analysis indicated that TLR9 was co-localized with EdU-positive deposits (Fig. 3e). TLR9 is activated by synthetic oligodeoxynucleotides (ODN1668) that contains unmethylated CpG[6], but it is inhibited by inhibitory ligands, such as ODN2088 (ref. 20), in which 'GCGTT' in ODN1668 is replaced with 'GCGGG'. ODN1668 induced increases in *Il1b* and *Il6* mRNA levels in wild-type isolated adult cardiomyocytes (data not shown). We, then, examined the effect of ODN2088 on carbonyl cyanide *m*-chlorophenyl hydrazone (CCCP) or isoproterenol-induced cell death using isolated adult cardiomyocytes to eliminate the contribution of immune cells[5]. CCCP, a protonophore, induces dissipation of mitochondrial membrane potential. Isoproterenol caused a loss of mitochondrial membrane potential in wild-type cardiomyocytes, as indicated by loss of tetramethylrhodamine ethyl ester signal (Supplementary Fig. 7a). Incubation with CCCP or isoproterenol induced conversion of LC3-I to LC3-II, an essential step during autophagosome formation, and treatment with the lysosomal inhibitor bafilomycin A1 led to an even larger increase of LC3-II in CCCP- or isoproterenol-treated cells than in control cells, indicating that CCCP or isoproterenol accelerated autophagic flux (Supplementary Fig. 7b). Isolated cardiomyocytes from *Dnase2a*$^{-/-}$ hearts were more susceptible than those from control hearts to CCCP or isoproterenol in the presence of

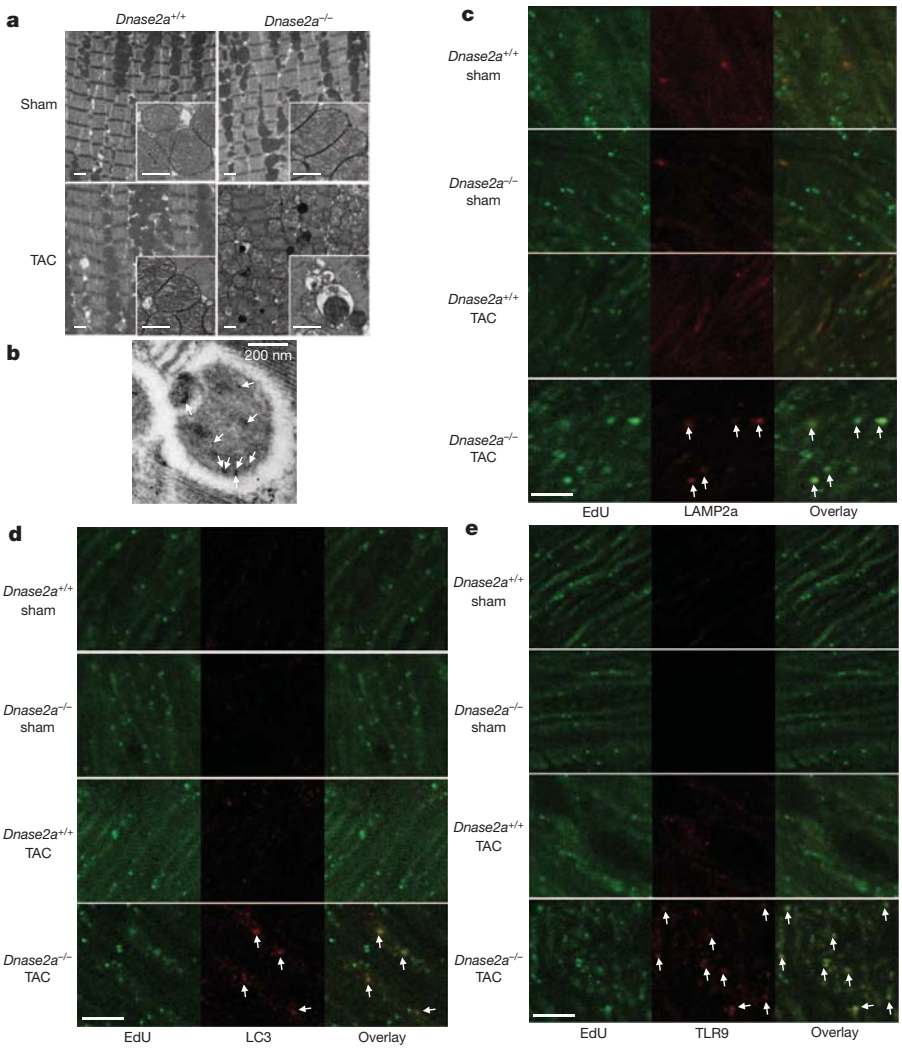


Figure 3 | Deposition of mitochondrial DNA in autolysosomes in pressure-overloaded *Dnase2a*$^{-/-}$ hearts. Mice were analysed 2 days after TAC (**a**–**e**). **a**, Electron microscopic analysis. Images of mitochondria at higher magnification are shown in subsets. Scale bar, 1 μm. **b**, Autolysosome after incubation with anti-DNA antibody and 10 nm gold staining. Scale bar, 200 nm. Arrows indicate labelled DNA. Double staining of heart sections with EdU (green) and anti-LAMP2a antibody (red) (**c**), EdU (green) and anti-LC3 antibody (red) (**d**) or EdU and anti-TLR9 antibody (red) (**e**). Arrows indicate EdU-positive and LAMP2a-, LC3- or TLR9-positive structures. Scale bar, 10 μm.

inactive control oligodeoxynucleotides (ODN2088 control) (Supplementary Fig. 7c–e). CCCP upregulated the mRNA expression levels of *Il1b* and *Il6* in *Dnase2a*$^{-/-}$ cardiomyocytes (Supplementary Fig. 7f). Incubation of *Dnase2a*$^{-/-}$ cardiomyocytes with ODN2088 attenuated the cell death and cytokine mRNA induction by CCCP treatment. Treatment of the *Dnase2a*$^{-/-}$ cardiomyocytes with 3-methyladenine, an autophagy inhibitor, and rapamycin, an autophagy inducer, inhibited and enhanced the induction of the cytokine mRNA by CCCP treatment, respectively (Supplementary Fig. 7g).

We next examined whether the inhibition of TLR9 can rescue the cardiac phenotypes in TAC-operated *Dnase2a*$^{-/-}$ mice. Administration of ODN2088 resulted in the improvement of survival 28 days after TAC (Fig. 4a). ODN2088 attenuated chamber dilation and cardiac dysfunction compared with the control oligodeoxynucleotides 4 days after TAC (Fig. 4b, c and Supplementary Fig. 8a). In addition, ODN2088 inhibited infiltration of CD68^{+} macrophages and Ly6G^{+} cells, fibrosis and upregulation of *Il6*, *Ifng*, *Nppa* and *Col1a2* mRNAs in TAC-operated *Dnase2a*$^{-/-}$ hearts (Fig. 4d and Supplementary Fig. 8b–e). ODN2088 prevented cardiac remodelling for a longer time (end-diastolic left ventricular internal dimension (LVIDd), in millimetres, 2.74 ± 0.03, 2.76 ± 0.03; end-systolic left ventricular internal dimension (LVIDs), in millimetres, 1.37 ± 0.03, 1.34 ± 0.05; left ventricular fractional shortening (LVFS) (%), 50.1 ± 0.7, 51.4 ± 1.5, before and 14 days after TAC, respectively, $n = 6$). Furthermore, ablation of *Tlr9* rescued the cardiac phenotypes in TAC-operated *Dnase2a*$^{-/-}$ mice (Supplementary Fig. 9).

To examine the significance of TLR9 signalling pathway in the genesis of heart failure, we subjected TLR9-deficient mice[6] to TAC. Ten weeks after TAC, TLR9-deficient mice showed smaller left ventricular dimensions, better cardiac function and less pulmonary congestion than in TAC-operated control mice (Fig. 4e, f and Supplementary Fig. 10a). The extent of fibrosis, the levels of *Nppa*, *Nppb* and *Col1a2* mRNA, infiltration of CD68^{+} macrophages were attenuated in TLR9-deficient mice (Supplementary Fig. 10b–e). We detected no significant differences in the cytokine mRNA levels between TAC-operated groups (Supplementary Fig. 10f). Furthermore, ODN2088 improved survival of wild-type mice in a more severe TAC model (Supplementary Fig. 10g). These data indicate that the TLR9 signalling pathway is involved in inflammatory responses in failing hearts in response to pressure overload and plays an important role in the pathogenesis of heart failure.

In this study, we showed that mitochondrial DNA that escapes from autophagy-mediated degradation cell-autonomously leads to TLR9-mediated inflammatory responses in cardiomyocytes, myocarditis and dilated cardiomyopathy. Immune responses are initiated and perpetuated by endogenous molecules released from necrotic cells, in addition to pathogen-associated molecular patterns expressed in invading microorganisms[21]. Cellular disruption by trauma releases mitochondrial molecules, including DNA, into circulation to cause systemic inflammation[19]. Depletion of autophagic proteins promotes cytosolic translocation of mitochondrial DNA and caspase-1-dependent cytokines mediated by the NALP3 inflammasome in response to

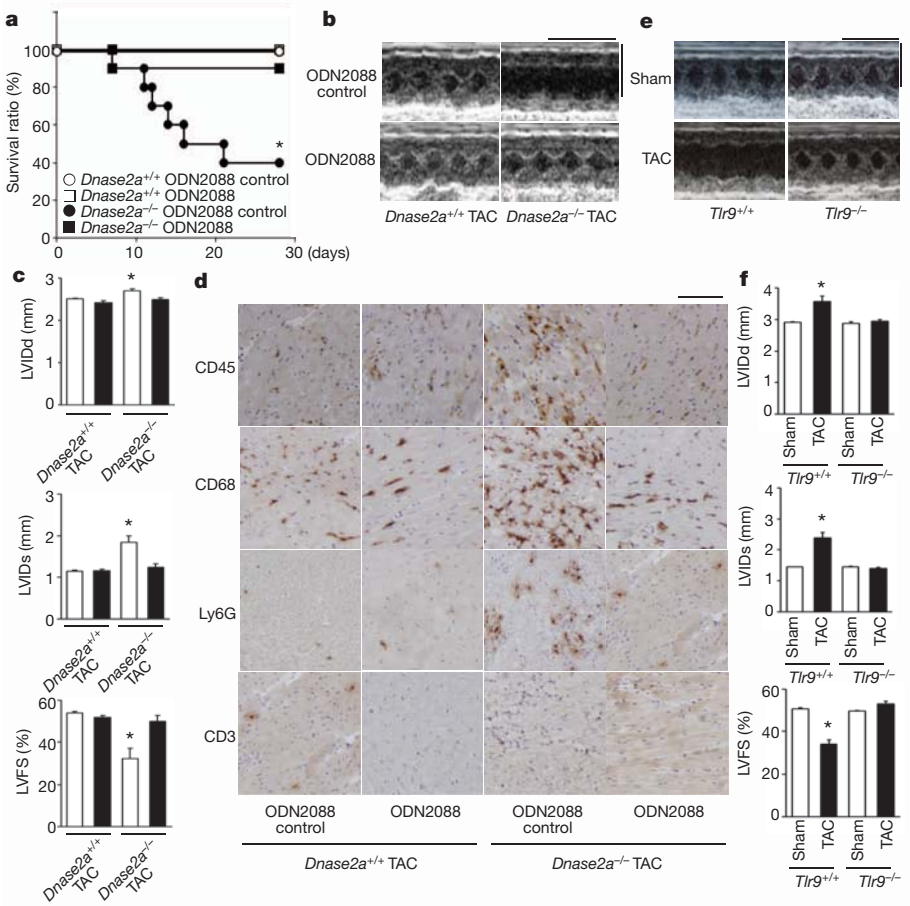


Figure 4 | Inhibition of TLR9 attenuated TAC-induced heart failure. **a**, Survival ratio of TAC-operated ODN-treated mice ($n = 6$–10 per group). **b–d**, Four days after TAC. **b**, Echocardiography. Scale bars, 0.2 s and 5 mm. **c**, Echocardiographic parameters. Open and closed bars represent ODN2088 control- and ODN2088-treated groups, respectively ($n = 5$–8 per group). **d**, Immunohistochemical analysis. Scale bar, 100 μm. TLR9-deficient mice were analysed 10 weeks after TAC (**e**, **f**). **e**, Scale bars, 0.2 s and 5 mm. **f**, Echocardiographic parameters ($n = 6$–10 per group). Data are mean ± s.e.m. *$P < 0.05$ versus all other groups.

lipopolysaccharide in macrophages[22]. We observed no significant difference in the amount of mitochondrial DNA in the blood between TAC-operated *Dnase2a*$^{-/-}$ and *Dnase2a*$^{+/+}$ mice (data not shown), excluding a possibility that circulating mitochondrial DNA is causing most of the inflammatory responses mediated by TLR9. The mechanisms presented here do not require release of mitochondrial DNA from cardiomyocytes into extracellular space.

Increased levels of circulating proinflammatory cytokines are associated with disease progression and adverse outcomes in patients with chronic heart failure[1]. Mitochondrial DNA plays an important role in inducing and maintaining inflammation in the heart. This mechanism might work in many chronic non-infectious inflammation-related diseases such as atherosclerosis, metabolic syndrome and diabetes mellitus.

METHODS SUMMARY

Animal study. The study was performed under the supervision of the Animal Research Committee of Osaka University and in accordance with the Japanese Act on Welfare and Management of Animals (No. 105). The 12- to 14-week-old mice were subjected to TAC[5,23] and severe TAC using 26- and 27-gauge needles for aortic constriction, respectively.

Biochemical assays. The DNase II activity was determined by the single radial enzyme-diffusion method[24]. The mRNA levels were determined by quantitative PCR with reverse transcription (RT–PCR)[5].

Histological analysis. The antibodies used were anti-mouse CD45 (ANASPEC), CD68 (Serotec), Ly6G/6C (BD Pharmingen), CD3 (Abcam), DNA (Abcam), LAMP2a (Zymed), LC3 (ref. 25) and TLR9 (Santa-Cruz). The *in situ* hybridization analysis was performed using DIG RNA Labelling Kit and DIG Nucleic Acid Detection Kit (Roche Diagnostics). Hearts were embedded in LR White resin for immunoelectron microscopy[26]. Heart sections were incubated in PicoGreen (Molecular Probes) for 1 h. Twenty-four hours before TAC, mice were injected intraperitoneally with 250 μg of EdU every 2 h five times, and EdU was detected with a Click-iT EdU Alexa Fluor 488 Imaging Kit (Invitrogen).

***In vitro and in vivo* rescue experiments with the TLR9 inhibitor.** Cardiomyocytes[5] were pre-treated with 1 μg ml^{-1} inhibitory CpG (ODN2088) or control (ODN2088 control) oligodeoxynucleotides for 5 h and incubated with 20 nM CCCP or 50 μM isoproterenol for 24 h (ref. 20). The cells were loaded with tetramethylrhodamine ethyl ester (Molecular Probe) at 10 nM for 30 min. The mice were injected intravenously with 500 μg of the oligodeoxynucleotides 2 h before and 2 and 4 days after TAC, and every 3 days thereafter.

Statistical analysis. Results are shown as the mean ± s.e.m. Paired data were evaluated using a Student's *t*-test. A one-way analysis of variance with the Bonferroni post hoc test was used for multiple comparisons. The Kaplan–Meier method with a log-rank test was used for survival analysis.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 26 June 2011; accepted 1 March 2012.
Published online 25 April 2012.

1. Mann, D. L. Inflammatory mediators and the failing heart: past, present, and the foreseeable future. *Circ. Res.* **91,** 988–998 (2002).
2. Pollack, Y., Kasir, J., Shemer, R., Metzger, S. & Szyf, M. Methylation pattern of mouse mitochondrial DNA. *Nucleic Acids Res.* **12,** 4811–4824 (1984).
3. Cardon, L., Burge, C., Clayton, D. A. & Karlin, S. Pervasive CpG suppression in animal mitochondrial genomes. *Proc. Natl Acad. Sci. USA* **91,** 3799–3803 (1994).
4. Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283,** 1476–1481 (1999).
5. Nakai, A. *et al.* The role of autophagy in cardiomyocytes in the basal state and in response to hemodynamic stress. *Nature Med.* **13,** 619–624 (2007).
6. Hemmi, H. *et al.* A Toll-like receptor recognizes bacterial DNA. *Nature* **408,** 740–745 (2000).
7. Taanman, J.-W. The mitochondrial genome: structure, transcription, translation and replication. *Biochim Biophys Acta Bioenerget.* **1410,** 103–123 (1999).
8. Collins, L., Hajizadeh, S., Holme, E., Jonsso, I. & Tarkowski, A. Endogenously oxidized mitochondrial DNA induces *in vivo* and *in vitro* inflammatory responses. *J. Leukoc. Biol.* **75,** 995–1000 (2004).
9. Mizushima, N., Levine, B., Cuervo, A. M. & Klionsky, D. J. Autophagy fights disease through cellular self-digestion. *Nature* **451,** 1069–1075 (2008).
10. Meerson, F., Zaletayeva, T., Lagutchev, S. & Pshennikova, M. Structure and mass of mitochondria in the process of compensatory hyperfunction and hypertrophy of the heart. *Exp. Cell Res.* **36,** 568–578 (1964).
11. Bugger, H. *et al.* Proteomic remodelling of mitochondrial oxidative pathways in pressure overload-induced heart failure. *Cardiovasc. Res.* **85,** 376–384 (2010).
12. Evans, C. J. & Aguilera, R. J. DNase II: genes, enzymes and function. *Gene* **322,** 1–15 (2003).
13. Kawane, K. *et al.* Chronic polyarthritis caused by mammalian DNA that escapes from degradation in macrophages. *Nature* **443,** 998–1002 (2006).
14. Ashley, N., Harris, D. & Poulton, J. Detection of mitochondrial DNA depletion in living human cells using PicoGreen staining. *Exp. Cell Res.* **303,** 432–446 (2005).
15. Kabeya, Y. *et al.* LC3, a mammalian homologue of yeast Apg8p, is localized in autophagosome membranes after processing. *EMBO J.* **19,** 5720–5728 (2000).
16. Yamaguchi, O. *et al.* Cardiac-specific disruption of the *c-raf-1* gene induces cardiac dysfunction and apoptosis. *J. Clin. Invest.* **114,** 937–943 (2004).
17. Lentz, S. I. *et al.* Mitochondrial DNA (mtDNA) biogenesis: visualization and duel incorporation of BrdU and EdU into newly synthesized mtDNA *in vitro*. *J. Histochem. Cytochem.* **58,** 207–218 (2010).
18. Takeuchi, O. & Akira, S. Pattern recognition receptors and inflammation. *Cell* **140,** 805–820 (2010).
19. Zhang, Q. *et al.* Circulating mitochondrial DAMPs cause inflammatory responses to injury. *Nature* **464,** 104–107 (2010).
20. Stunz, L. *et al.* Inhibitory oligonucleotides specifically block effects of stimulatory CpG oligonucleotides in B cells. *Eur. J. Immunol.* **32,** 1212–1222 (2002).
21. Bianchi, M. E. DAMPs, PAMPs and alarmins: all we need to know about danger. *J. Leukoc. Biol.* **81,** 1–5 (2007).
22. Nakahira, K. *et al.* Autophagy proteins regulate innate immune responses by inhibiting the release of mitochondrial DNA mediated by the NALP3 inflammasome. *Nature Immunol.* **12,** 222–230 (2011).
23. Yamaguchi, O. *et al.* Targeted deletion of apoptosis signal-regulating kinase 1 attenuates left ventricular remodeling. *Proc. Natl Acad. Sci. USA* **100,** 15883–15888 (2003).
24. Koizumi, T. Deoxyribonuclease II (DNase II) activity in mouse tissues and body fluids. *Exp. Anim.* **44,** 169–171 (1995).
25. Lu, Z. *et al.* Participation of autophagy in the degeneration process of rat hepatocytes after transplantation following prolonged cold preservation. *Arch. Histol. Cytol.* **68,** 71–80 (2005).
26. Mosgoller, W. *et al.* Distribution of DNA in human Sertoli cell nucleoli. *J. Histochem. Cytochem.* **41,** 1487–1493 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Nagata and K. Kawane, Kyoto University, for discussions and a gift of *Dnase2a*$^{flox/flox}$ mice, and Y. Uchiyama, Juntendo University, for anti-LC3 antibody. We also thank K. Takada for technical assistance. This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology in Japan and research grants from Mitsubishi Pharma Research Foundation and the British Heart Foundation (CH/11/3/29051, RG/11/12/29052).

Author Contributions S.A. and I.K. provided intellectual input; K.O. was responsible for the overall study design and writing the manuscript. The other authors performed experiments and analysed data. All authors contributed to the discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.O. (kinya.otsu@kcl.ac.uk).

METHODS

Animal study. The study was performed under the supervision of the Animal Research Committee of Osaka University and in accordance with the Japanese Act on Welfare and Management of Animals (No. 105).

We crossed mice bearing a *Dnase2a*^{flox} allele[13] with transgenic mice expressing *Cre* recombinase under the control of the α-myosin heavy chain promoter (α-MyHC)[16], to produce cardiac-specific DNase II-deficient mice, *Dnase2a*$^{flox/flox}$;α-MyHC-*Cre*$^{+}$ (*Dnase2a*$^{-/-}$). To generate double-knockout mice of *Dnase2a* and *Tlr9*, we crossed *Dnase2a*$^{-/-}$ mice with *Tlr9*$^{-/-}$ mice[6].

The 12- to 14-week-old male mice were subjected to TAC[5,23] and severe TAC using 26- and 27-gauge needles for aortic constriction, respectively. Non-invasive measurements of blood pressure were performed on mice anaesthetized with 2.5% avertin using a blood pressure monitor for rats and mice Model MK-2000 (Muromachi Kikai) according to the manufacturer's instructions[5,23]. To perform echocardiography on awakened mice, ultrasonography (SONOS-5500, equipped with a 15 MHz linear transducer, Philips Medical Systems) was used. The heart was imaged in the two-dimensional parasternal short-axis view, and an M-mode echocardiogram of the midventricle was recorded at the level of the papillary muscles. Heart rate, intraventricular septum and posterior wall thickness, and end-diastolic and end-systolic internal dimensions of the left ventricle were obtained from the M-mode image.

Measurement of DNase II activity. The DNase II activity was determined using the single radial enzyme-diffusion method[24]. The heart homogenates were applied to the cylindrical wells (radius, 1.5 mm) punched in 1% (w/v) agarose gel, containing 0.05 mg ml^{-1} salmon sperm DNA (Type III), 5 μg ml^{-1} ethidium bromide, 0.5 M sodium acetate buffer (pH 4.7) and 10 mM EDTA. After incubation for 48 h at 37 °C, the radius of the dark circle was measured under an ultraviolet transilluminator at 312 nm. DNase II activities for the samples were determined using a standard curve constructed from the serial dilution of porcine DNase II (Sigma).

Quantitative RT–PCR. Total RNA was isolated from the left ventricle or cultured cardiomyocytes for analysis using the TRIzol reagent (Invitrogen Life Technologies). The mRNA levels were determined by quantitative RT–PCR[5]. For reverse transcription and amplification, we used the TaqMan Reverse Transcription Reagents (Applied Biosystems) and Platinum Quantitative PCR SuperMix-UDG (Invitrogen Life Technologies), respectively. The PCR primers and probes were obtained from Applied Biosystems. The primers used were as follows: *Nppa* assay identity, mm01255747_g1; *Nppb* assay identity, mm00435304_g1; *Col1a2* assay identity, Mm01165187_m1; *Gapdh* assay identity, 4352339E; *Il6* assay identity, Mm99999064_m1; *Il1b* assay identity, Mm01336189_m1; *Ifnb1* assay identity, Mm00439546_s1; *Ifng* assay identity, Mm99999071_m1; *Tnf* assay identity, Mm00443260_g1; and *Dnase2a* assay identity, Mm00438463_m1. We constructed quantitative PCR standard curves using the corresponding complementary DNA, and all data were normalized to *Gapdh* mRNA content.

Histological analysis. Heart samples were excised and immediately fixed in buffered 4% paraformaldehyde, embedded in paraffin and cut into 5 μm sections. Haematoxylin and eosin or AZAN-Mallory staining was performed on serial sections[5,23]. Myocyte cross-sectional area was measured by tracing the outline of 100–200 myocytes in each section[5,23]. For immunohistochemical analysis, frozen heart sections (5 μm) were fixed in buffered 4% paraformaldehyde. The antibodies used were anti-mouse CD45 (ANASPEC), CD68 (Serotec), Ly6G/6C (BD Pharmingen), CD3 (Abcam), LAMP2a (Zymed), LC3 (ref. 25) and TLR9 (Santa-Cruz). For *in situ* hybridization analysis, the mouse IL-6 (1-636) and IL-1β (1-810) RNA probes were labelled using a DIG RNA Labelling Kit and detected using a DIG Nucleic Acid Detection Kit (Roche Diagnostics). For immunoelectron microscopy, frozen heart tissue was embedded in LR White resin and the deposited DNA was detected using anti-DNA antibody (Abcam) and immunogold conjugated anti-mouse IgG (British Biocell International)[26]. For DNA detection, heart sections were incubated in PicoGreen (Molecular Probes) for 1 h. We used EdU to detect mitochondrial DNA in the heart section. Twenty-four hours before TAC, mice were injected intraperitoneally with 250 μg of EdU every 2 h five times, and EdU was detected using a Click-iT EdU Alexa Fluor 488 Imaging Kit (Invitrogen).

***In vitro and in vivo* rescue experiments with the TLR9 inhibitor.** Adult mouse cardiomyocytes were isolated from 12- to 14-week-old male mouse hearts as we previously described[5]. Cardiomyocytes were pre-treated with 1 μg ml^{-1} inhibitory CpG oligodeoxynucleotides (ODN2088) (Operon) (5′-TCCTGGCGGGGAAGT-3′) or control oligodeoxynucleotides (ODN2088 control) (5′-TCCTGAGCTTGAAGT-3′) for 5 h and incubated with 20 nM CCCP or 50 μM isoproterenol for 24 h (ref. 20). Cell death was estimated by Trypan blue staining[5]. To monitor mitochondrial membrane potential ($\Delta\psi$), the cells were loaded with tetramethylrhodamine ethyl ester (Molecular Probes) at 10 nM for 30 min before observation. In *in vivo* study, the mice were injected intravenously with 500 μg of the oligodeoxynucleotides 2 h before and 2 days after TAC, and they were analysed 4 days after TAC. To estimate survival, the mice received additional administration of the oligodeoxynucleotides 4 days after TAC and every 3 days thereafter.

Statistical analysis. Results are shown as the mean ± s.e.m. Paired data were evaluated using a Student's *t*-test. A one-way analysis of variance with the Bonferroni post hoc test was used for multiple comparisons. The Kaplan–Meier method with a log-rank test was used for survival analysis.

LETTER

doi:10.1038/nature10996

Cell attachment protein VP8* of a human rotavirus specifically interacts with A-type histo-blood group antigen

Liya Hu[1], Sue E. Crawford[2], Rita Czako[2], Nicolas W. Cortes-Penfield[2], David F. Smith[3], Jacques Le Pendu[4,5,6], Mary K. Estes[2] & B. V. Venkataram Prasad[1,2]

As with many other viruses, the initial cell attachment of rotaviruses, which are the major causative agent of infantile gastroenteritis, is mediated by interactions with specific cellular glycans[1–4]. The distally located VP8* domain of the rotavirus spike protein VP4 (ref. 5) mediates such interactions. The existing paradigm is that 'sialidase-sensitive' animal rotavirus strains bind to glycans with terminal sialic acid (Sia), whereas 'sialidase-insensitive' human rotavirus strains bind to glycans with internal Sia such as GM1 (ref. 3). Although the involvement of Sia in the animal strains is firmly supported by crystallographic studies[1,3,6,7], it is not yet known how VP8* of human rotaviruses interacts with Sia and whether their cell attachment necessarily involves sialoglycans. Here we show that VP8* of a human rotavirus strain specifically recognizes A-type histo-blood group antigen (HBGA) using a glycan array screen comprised of 511 glycans, and that virus infectivity in HT-29 cells is abrogated by anti-A-type antibodies as well as significantly enhanced in Chinese hamster ovary cells genetically modified to express the A-type HBGA, providing a novel paradigm for initial cell attachment of human rotavirus. HBGAs are genetically determined glycoconjugates present in mucosal secretions, epithelia and on red blood cells[8], and are recognized as susceptibility and cell attachment factors for gastric pathogens like *Helicobacter pylori*[9] and noroviruses[10]. Our crystallographic studies show that the A-type HBGA binds to the human rotavirus VP8* at the same location as the Sia in the VP8* of animal rotavirus, and suggest how subtle changes within the same structural framework allow for such receptor switching. These results raise the possibility that host susceptibility to specific human rotavirus strains and pathogenesis are influenced by genetically controlled expression of different HBGAs among the world's population.

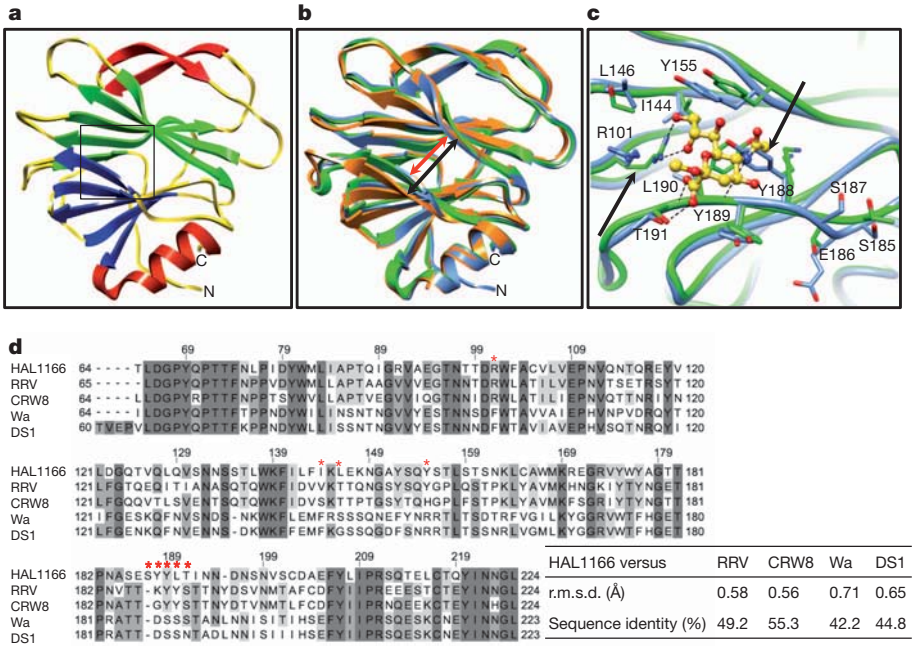
Rotaviruses are classified on the basis of the neutralization specificity of the outer capsid proteins VP7 and VP4 into G (VP7) and P (VP4) genotypes following a dual nomenclature system similar to influenza viruses[11]. The crystallographic structures of VP8* from two sialidase-insensitive human strains, representing P[8] (Wa)[1] and P[4] (DS1)[12], from two sialidase-sensitive animal strains, representing P[3] (RRV)[6,7]and P[7] (CRW-8)[1], and the structures of two animal VP8* with bound Sia[1,6,12] have been previously reported. NMR, cell-binding and neutralization assays showed that the sialidase-insensitive P[8] Wa strain binds to gangliosides such as GM1 using internal Sia[3]. These studies suggested that whereas the sialidase-sensitive strains recognize glycans with terminal Sia such as GD1a, the sialidase-insensitive rotavirus strains bind to gangliosides such as GM1 with an internal Sia moiety, and gave rise to the notion that Sia is the key determinant for host-cell recognition in rotaviruses. Our goal was to determine whether all sialidase-insensitive human rotavirus genotypes recognize gangliosides with an internal Sia moiety for initial cell attachment, or whether they recognize different glycans in a genotype-dependent manner. VP8* (amino acids 64–224), cloned from a human rotavirus strain (HAL1166) first isolated from a child in Finland[13], was expressed in *Escherichia coli*, purified to homogeneity and crystallized for structural analysis. The sialidase-insensitive HAL1166 strain, phylogenetically and serologically belongs to G8P[14] genotype[14]. Although not as prevalent as the P[4] and P[8] genotypes, the P[14] genotypes are being increasingly documented by global rotavirus surveillance[15–17], and P[14] human rotaviruses are thought to be able to jump from animal to human hosts[17].

The structure of the HAL1166 VP8* determined to 1.5 Å resolution shows the characteristic galectin-like fold with two twisted β-sheets separated by a shallow cleft as observed in the VP8* structures from other rotavirus strains (Fig. 1a). The structure of P[14] VP8* superimposes well with all of the VP8* structures previously determined. One significant difference between these structures is in the width of the cleft separating the two twisted β-barrel sheets (Fig. 1b). In P[14] VP8*, it is narrower than the cleft in the VP8* of the other two human strains, similar to that in the VP8* of the animal strains. In the animal VP8* structures, Sia binds near the cleft (Fig. 1c). Although the cleft in the P[14] VP8* structure is of similar dimensions as in the animal VP8* structures, the structural features in this region of P[14] VP8* is not compatible with Sia binding. In addition to changes in the amino acid residues (Fig. 1d) and side-chain orientations, the positioning of the amino acid residues is slightly shifted in this region because of an insertion (amino acid 187). The side-chain of Y188 is oriented in such a way that it would cause steric hindrance to Sia binding (Fig. 1c). Furthermore, the P[14] VP8* structure with a narrower cleft and several amino acid changes (Fig. 1d) is not compatible with binding of GM1 as suggested, based on computer modelling, for VP8* of the other human strains with a wider cleft[3].

These observations prompted us to undertake a high throughput screening of a glycan array comprised of 511 different glycans, including several glycans with terminal or internal Sia. Such screening, which has been used to identify cellular glycans for a variety of pathogens including bacterial toxins[18], influenza viruses[19] and polyomavirus[20], unambiguously showed specific binding to glycans with a terminal GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAc sequence (Fuc, fucose; Gal, galactose; Glc, glucose; NAc, *N*-acetyl, which is a characteristic of A-type HBGA (Supplementary Table 1, Supplementary Figs 1 and 2). None of the sialylated glycans, with either internal or terminal Sia, showed significant binding (Supplementary Table 1, blue).

To understand the structural interactions between P[14] VP8* and A-type HBGA, we co-crystallized VP8* with tri- and tetrasaccharides that correspond to the terminal structure in the A-type HBGA. The structure of the complex, determined to a similar resolution of 1.5 Å as the unliganded structure, clearly showed density for the bound ligand

[1]Verna and Marrs McLean, Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, Texas 77030, USA. [3]Department of Biochemistry, Emory University School of Medicine, Atlanta, Georgia 30322, USA. [4]NSERM, UMR892, Université de Nantes, 44007 Nantes, France. [5]CNRS, UMR 6299, Université de Nantes, 44007 Nantes, France. [6]Université de Nantes, 44007 Nantes, France.



HAL1166 versus	RRV	CRW8	Wa	DS1
r.m.s.d. (Å)	0.58	0.56	0.71	0.65
Sequence identity (%)	49.2	55.3	42.2	44.8

Figure 1 | VP8* structure of HAL1166 P[14] human rotavirus strain and structural comparison with other VP8* structures. **a**, Cartoon representation of the P[14] VP8* structure showing a galectin-like fold with the characteristic two twisted β-sheets (in blue and green) separated by a cleft. The Sia binding site as observed in the animal VP8* is shown by a box. The β-ribbon and the carboxy-terminal α-helix of the structure are shown in red. The amino and carboxy termini are denoted. **b**, Structural alignment of P[14] VP8* structure (blue) with VP8* structure of sialidase-sensitive animal rotavirus RRV strain (green, PDB ID 1KQR) and VP8* of sialidase-insensitive human rotavirus strain Wa (orange, PDB ID 2DWR) shown in the same orientation as in **a**. The width of the cleft in the P[14] VP8* is narrower (red arrow) than in the Wa VP8* structure (black arrow). **c**, P[14] VP8* shows changes in the amino acid composition and side-chain orientations in the region corresponding to Sia binding site of RRV VP8*. The amino acid residues interacting with Sia in the RRV VP8* structure are shown as green sticks, and bound Sia is shown as yellow sticks with oxygen atoms in red. The residues in this region of the P[14] VP8* structure are shown as blue sticks. The residue numbering corresponds to HAL1166 VP8*. The key amino acid changes in this region between P[14] and RRV VP8* are indicated by black arrows. Noticeable is how Y188 in the P[14] VP8* structure causes steric hindrance if Sia were to bind in this region. Also noticeable is the change in side-chain orientation of the conserved R101. **d**, Alignment of HAL1166 VP8* with other VP8* sequences from sialidase-sensitive animal (RRV, CRW-8) strains, and sialidase-insensitive human rotavirus (Wa and DS1) strains. The residues that interact with Sia in the animal VP8* as shown in **c** are indicated by red stars on the top of the sequence. Highly conserved (>80%), and moderately conserved (>60%) regions are coloured in dark and lighter grey, respectively. The root mean square deviation (r.m.s.d.) of the matching Cα atoms between the P[14] VP8* and other VP8* structures along with percentage of sequence identity are shown in the table on the right.

(Fig. 2a). The HBGA binds near the cleft region at the same location as the Sia in the VP8* structures of the animal rotavirus strains (Fig. 2b and Supplementary Fig. 3). The binding of HBGA, which does not cause any conformational changes in the VP8*, involves a network of both direct and solvent-mediated hydrogen bond interactions, in addition to several stabilizing hydrophobic contacts (Supplementary Fig. 4). The terminal GalNAc and Gal of the HBGA participate in all of the interactions (Fig. 2c). The GalNAc participates in direct hydrogen-bonding interactions involving the side chains of R101 and T191, and hydrophobic interactions involving L190 and T191, whereas Gal participates in hydrogen-bonding interactions with the main chain carbonyl groups of Y189 and S187, and hydrophobic interactions with Y189 and Y188. The proximal sugar moieties project out from the surface of VP8* without making any direct contacts with VP8* (Fig. 2 and Supplementary Fig. 3).

The structure of P[14] VP8* with HBGA shows how subtle amino acid changes result in altered ligand specificity. The only positionally conserved common residue between P[14] and animal VP8* that participates in the ligand interactions is R101. Although the position matches well, its side-chain orientation differs significantly between the two structures (Fig. 1c). Without this change, the side-chain of R101 in the P[14] VP8* structure would cause steric hindrance to the GalNAc moiety of the HBGA. Most other residues that participate in the ligand interactions are upstream of position 187 in the VP8* sequence, where an insertion occurs in the P[14] VP8*. This insertion along with sequence variation alters the configuration of the binding site in the P[14] VP8* to allow specific interaction with HBGA. Insertion at position 187 causes a localized change which makes the side-chain of Y188 clash with Sia when placed in the structure of P[14] VP8* (Fig. 1c). Following this residue, the polypeptide chain reverts back to the same course as in the animal VP8* structures such that Y189 and T191, which interact with HBGA, are now positionally equivalent to Y188 and S190 which interact with Sia in the animal VP8* (Figs 1c and 2c).

The remarkable overlap of the HBGA binding site in the P[14] VP8* with that of the Sia in the animal VP8* structure strongly suggests that A-type HBGA is a cell attachment factor for P[14] rotavirus strains. To examine the biologic relevance of HBGA binding to P[14] VP8*, virus infectivity assays were performed. Using intestinal HT29 cells isolated from a type A individual, dose-dependent abrogation of HAL1166 infectivity was observed, with a greater than 75% reduction at the highest concentration of anti-A-type HBGA antibody compared to an isotype control antibody (Fig. 3a, b). In contrast, this antibody did not inhibit the sialidase-sensitive SA11 rotavirus strain (Fig. 3a). To ascertain further the specificity of the HAL1166 to A-type HBGA, infectivity assays were performed using parental Chinese hamster ovary (CHO) cells, which do not express any HBGA, and genetically-engineered CHO cells expressing either A- or H-type HBGA. CHO cells expressing type A HBGA showed a large increase in infectivity with HAL1166, but not SA11, compared to parental CHO cells or those expressing type H alone (Fig. 3c). Similarly, low infectivity was observed in Caco-2 cells, isolated from a blood type O individual, compared to HT29 cells that express type A HBGA (data not shown). Specificity to A-type HBGA was further confirmed by performing

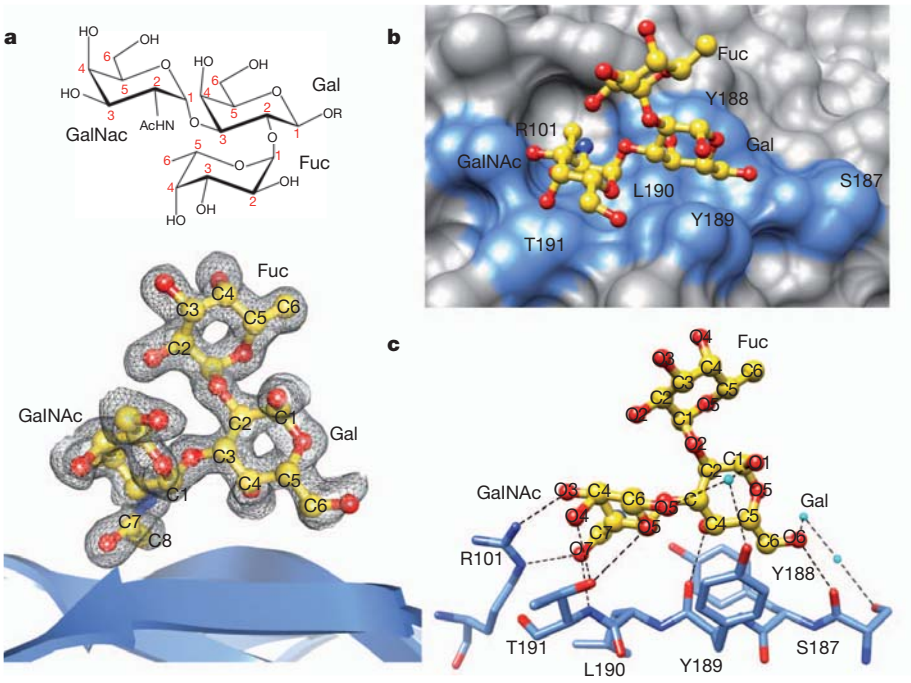


Figure 2 | Structural analysis of P[14] VP8*–A-type HBGA interactions. **a**, The chemical structure of the A-type trisaccharide (above) and simulated annealing omit difference map (below), contoured at 3σ level, showing the binding of A-type trisaccharide to P[14] VP8*. Bound A-type trisaccharide (GalNAc, N-acetylgalactosamine; Gal, galactose; Fuc, fucose) is shown in a ball-and-stick representation (yellow) inside the map with its carbon atoms numbered following the standard convention. The nitrogen and the oxygen atoms in the trisaccharide are coloured in blue and red, respectively. **b**, Surface representation of the P[14] VP8* structure (grey) with the bound A-trisaccharide shown in stick representation (with the same colour scheme as in **a**). The acetamido group of GalNAc inserts into a well-defined pocket in the VP8* structure. The amino acid residues in the P[14] VP8* which participate in hydrogen bond and hydrophobic interactions with the trisaccharide are indicated in blue. **c**, Network of hydrogen bond interactions (dashed lines) between the VP8* residues (light blue) and A-type trisaccharide (coloured as in **a**). Participating water molecules are shown as small spheres (cyan). More detailed interactions between VP8* and the ligand are given in Supplementary Fig. 4. The terminal two saccharide moieties of the A-type tetrasaccharide (GalNAcα1-3(Fucα1-2)Galβ1-4GlcNAc) also show similar interactions with the VP8* (Supplementary Fig. 3).

haemagglutination assays using P[14] glutathione-*S*-transferase (GST)–VP8*. Type A blood cells, but not type O or B, were haemagglutinated by soluble VP8* (Supplementary Fig. 5).

This is the first study describing the structural interactions between a human VP8* and a cellular glycan. Importantly, it shows that binding of sialylated glycans is not obligatory among sialidase-insensitive human rotavirus strains. Our finding that P[14] VP8* specifically recognizes HBGA raises important questions such as whether other human rotavirus strains interact with similar or other HBGAs in a serotype-dependent manner like in human noroviruses[10], and whether genetically controlled differential expression of HBGAs among world's population plays a role in susceptibility to human rotaviruses. In a recently published paper, G8P[14] rotavirus was identified in the stool samples from two adults with diarrhoea, who lived in the same geographical area in Denmark[21]. The blood type of one of these patients and of another patient infected with a G6P[14] virus was type A (B.

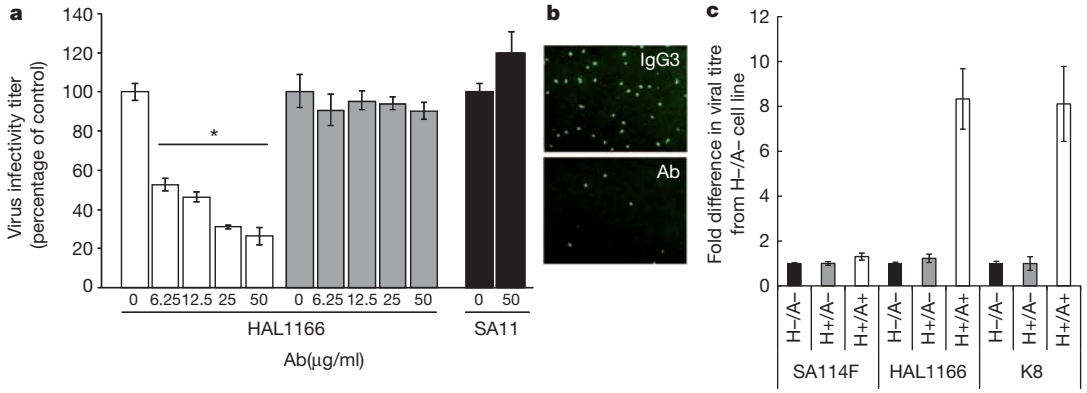


Figure 3 | HAL1166 rotavirus specifically recognizes A-type HBGA. **a**, Dose-dependent inhibition of HAL1166 infection in HT29 cells by anti-A-type antibody (Ab, white bars for HAL1166, and black bars for SA11). Isotype control IgG3 did not inhibit HAL1166 infectivity (grey bars). Error bars (also in Fig. 3c) represent standard deviation and the *P* values were determined by Student's *t*-test, $n = 3$. *All concentrations of anti-A-type antibody reduced infectivity compared to control with $P < 0.05$. **b**, Representative immunofluorescence microscopy images of HT29 cells infected with HAL1166 rotavirus in the presence of 50 μg ml^{-1} of IgG3 (top) and anti-A-type antibody (bottom). **c**, Infectivity of SA11 P[1], HAL1166 P[14] and K8 P[9] rotavirus strains in the parental CHO cells (H–/A–), the single transfectant with the Fut2 enzyme (H+/A–), and the double transfectant with both Fut2 and A type histo-blood group glycosyltranferase (H+/A+). The fold difference in infectivity was determined compared to parental cells. For HAL1166 and K8 human rotaviruses, the increase in infectivity in CHO (A+/H+) cells was compared to parental CHO, and CHO (H+/A–), the *P* values were <0.01.

Böttiger, personal communication). Although this is a small sample size, these findings warrant further epidemiological studies to determine whether HBGA is a susceptibility factor for rotaviruses. Based on sequence comparisons, our prediction that VP8* of the K8 human rotavirus strain (P[9] genotype) would also recognize A-type HBGA (Supplementary Fig. 6) is firmly supported by infectivity assays with parental and derivative CHO cells (Fig. 3c).

The double-stranded RNA rotaviruses, accounting for approximately 500,000 child deaths annually worldwide[22], have enormous genetic and strain diversity. In addition to point mutations and gene rearrangements, genetic reassortment between co-circulating strains, similar to influenza viruses, contribute to the expanding diversity of rotaviruses[23,24]. Current evidence indicates that many of the human rotavirus strains, including the P[14] HAL1166 strain[17], originated from animal reservoirs through reassortment and inter-species transmission[23,24]. Although effective vaccines are currently available, whether they will remain effective with such expanding virus diversity is an open question[25,26]. Discovery that a human rotavirus strain with host-switching capabilities binds to a non-sialylated but novel glycan receptor opens new approaches to better understand the molecular basis of critical human rotavirus–host interactions, which probably influences host specificity, cell specificity, pathogenesis and virus evolution.

METHODS SUMMARY

Expression, purification, and crystallization of P[14] VP8* and its complex with A-type oligosaccharides, and structure determination, using RRV VP8* structure (PDB ID: 1KQR) as a molecular replacement model, and refinement was carried out as described in the Methods section. Diffraction data were collected at Baylor College of Medicine using a Rigaku FR-E+ SuperBright rotating anode. See Supplementary Table 2 for data collection and refinement statistics. The carbohydrate-binding specificity of P[14] VP8* was investigated using glycan array v4.2 with 511 glycans in replicates of six (Consortium for Functional Glycomics Protein-Glycan Interaction Core (H) (http://www.functionalglycomics.com). GST-tagged VP8* bound on the glycan array was detected using a fluorescent-labelled anti-GST monoclonal antibody. Infectivity assays were performed as previously described[3,27].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 21 June 2011; accepted 29 February 2012.
Published online 15 April 2012.

1. Blanchard, H., Yu, X., Coulson, B. S. & von Itzstein, M. Insight into host cell carbohydrate-recognition by human and porcine rotavirus from crystal structures of the virion spike associated carbohydrate-binding domain (VP8*). *J. Mol. Biol.* **367,** 1215–1226 (2007).
2. Dormitzer, P. R. *et al.* Specificity and affinity of sialic acid binding by the rhesus rotavirus VP8* core. *J. Virol.* **76,** 10512–10517 (2002).
3. Haselhorst, T. *et al.* Sialic acid dependence in rotavirus host cell invasion. *Nature Chem. Biol.* **5,** 91–93 (2009).
4. Lopez, S. & Arias, C. F. Early steps in rotavirus cell entry. *Curr. Top. Microbiol. Immunol.* **309,** 39–66 (2006).
5. Settembre, E. C., Chen, J. Z., Dormitzer, P. R., Grigorieff, N. & Harrison, S. C. Atomic model of an infectious rotavirus particle. *EMBO J.* **30,** 408–416 (2011).
6. Dormitzer, P. R., Sun, Z. Y., Wagner, G. & Harrison, S. C. The rhesus rotavirus VP4 sialic acid binding domain has a galectin fold with a novel carbohydrate binding site. *EMBO J.* **21,** 885–897 (2002).
7. Kraschnefski, M. J. *et al.* Effects on sialic acid recognition of amino acid mutations in the carbohydrate-binding cleft of the rotavirus spike protein. *Glycobiology* **19,** 194–200 (2009).
8. Marionneau, S. *et al.* ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. *Biochimie* **83,** 565–573 (2001).
9. Ilver, D. *et al. Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* **279,** 373–377 (1998).
10. Glass, R. I., Parashar, U. D. & Estes, M. K. Norovirus gastroenteritis. *N. Engl. J. Med.* **361,** 1776–1785 (2009).
11. Matthijnssens, J. *et al.* Uniformity of rotavirus strain nomenclature proposed by the Rotavirus Classification Working Group (RCWG). *Arch. Virol.* **156,** 1397–1413 (2011).
12. Monnier, N. *et al.* High-resolution molecular and antigen structure of the VP8* core of a sialic acid-independent human rotavirus strain. *J. Virol.* **80,** 1513–1523 (2006).
13. Gerna, G. *et al.* Identification of a new VP4 serotype of human rotaviruses. *Virology* **200,** 66–71 (1994).
14. Ciarlet, M. & Estes, M. K. Human and most animal rotavirus strains do not require the presence of sialic acid on the cell surface for efficient infectivity. *J. Gen. Virol.* **80,** 943–948 (1999).
15. Chitambar, S. D., Arora, R., Kolpe, A. B., Yadav, M. M. & Raut, C. G. Molecular characterization of unusual bovine group A rotavirus G8P[14] strains identified in western India: emergence of P[14] genotype. *Vet. Microbiol.* **148,** 384–388 (2011).
16. Fukai, K., Saito, T., Inoue, K. & Sato, M. Molecular characterization of novel P[14],G8 bovine group A rotavirus, Sun9, isolated in Japan. *Virus Res.* **105,** 101–106 (2004).
17. Matthijnssens, J. *et al.* Are human P[14] rotavirus strains the result of interspecies transmissions from sheep or other ungulates that belong to the mammalian order Artiodactyla? *J. Virol.* **83,** 2917–2929 (2009).
18. Byres, E. *et al.* Incorporation of a non-human glycan mediates human susceptibility to a bacterial toxin. *Nature* **456,** 648–652 (2008).
19. Stevens, J. *et al.* Glycan microarray analysis of the hemagglutinins from modern and pandemic influenza viruses reveals different receptor specificities. *J. Mol. Biol.* **355,** 1143–1155 (2006).
20. Neu, U. *et al.* Structure-function analysis of the human JC polyomavirus establishes the LSTc pentasaccharide as a functional receptor motif. *Cell Host Microbe* **8,** 309–319 (2010).
21. Midgley, S. E., Hjulsager, C. K., Larsen, L. E., Falkenhorst, G. & Bottiger, B. Suspected zoonotic transmission of rotavirus group A in Danish adults. *Epidemiol. Infect.*. doi:10.1017/S0950268811001981 (27 September 2011).
22. Parashar, U. D., Gibson, C. J., Bresse, J. S. & Glass, R. I. Rotavirus and severe childhood diarrhea. *Emerg. Infect. Dis.* **12,** 304–306 (2006).
23. Estes, M. K. & Kapikian, A. Z. in *Fields Virology* Vol. 2 (eds Knipe, D. M. & Howley, P. M.) 1917–1974 (Lippincott Williams & Wilkins, 2007).
24. Gray, J. & Iturriza-Gomara, M. Rotaviruses. *Methods Mol. Biol.* **665,** 325–355 (2011).
25. Angel, J., Franco, M. A. & Greenberg, H. B. Rotavirus vaccines: recent developments and future considerations. *Nature Rev. Microbiol.* **5,** 529–539 (2007).
26. Gentsch, J. R. *et al.* Serotype diversity and reassortment between human and animal rotavirus strains: implications for rotavirus vaccine programs. *J. Infect. Dis.* **192** (Suppl 1), S146–S159 (2005).
27. Guillon, P. *et al.* Inhibition of the interaction between the SARS-CoV spike protein and its cellular receptor by anti-histo-blood group antibodies. *Glycobiology* **18,** 1085–1093 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge the support from NIH grants AI36040 (to B.V.V.P.), AI 080656 and P30 DK56338 (to M.K.E.), GM62116 (to the Consortium for Functional Glycomics), and the Robert Welch foundation (Q1279) to B.V.V.P. We thank R. Atmar and S. Shanker for helpful discussions and BCM X-ray core facility for data collection.

Author Contributions L.H. carried out expression, purification, crystallization, diffraction data collection and structure determination. L.H., S.E.C., R.C. and N.W.C.-P. contributed to virus infectivity assays in HT29, CHO cells and haemagglutination assays and data analyses. D.F.S. contributed to glycan array experiments and analysis. J.L.P. provided parental and genetically modified CHO cells and advice. M.K.E. provided supervision and advice on cell infectivity assays and analysis. L.H. and B.V.V.P. analysed and interpreted the structural data. B.V.V.P. contributed to the overall direction of the project and wrote the manuscript with input from other authors.

Author Information: The coordinates and structure factors for the P[14] VP8* structures are deposited in the Protein Data Bank under accession numbers 4DRR (apo), 4DRV (with A-type trisaccharide) and 4DS0 (with A-type tetrasaccharide). Raw glycan array data are available at http://www.functionalglycomics.org/glycomics/publicdata/selectedScreens.jsp. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.V.V.P. (vprasad@bcm.edu).

METHODS

Protein expression and purification. VP8* (amino acids 64–224) of HAL1166 rotavirus strain (P[14] genotype) was cloned into an expression vector pGEX-2T (GE healthcare) with an N-terminal GST tag and a thrombin cleavage site. The recombinant GST-VP8* was expressed in *E. coli* BL21 (DE3) (Novagen) and purified by Glutathione Sepharose 4 Fast Flow (GE healthcare). The GST tag was cleaved by using thrombin before rebinding the protein mixtures onto a Glutathione Sepharose column to remove the GST, leaving Gly-Ser at the N terminus. The VP8* was then filtered and further purified by size-exclusion chromatography on a Superdex-75 (GE healthcare) column with 10 mM Tris, pH 7.4, 100 mM NaCl, 1 mM dithiothreitol (DTT). The concentration of the purified VP8* was determined by measuring absorbance at 280 nm and using an absorption coefficient of 43,010 M^{-1} cm^{-1} calculated using Vector NTI 11 software (Invitrogen).

Crystallization. Crystallization conditions for P[14] VP8* (13.5 mg ml^{-1}) were screened by hanging-drop vapour diffusion using the Mosquito crystallization robot (TTP LabTech) and visualized using Rock Imager (Formulatrix) at 20 °C. The crystals from one of the conditions (30% PEG 1500, sodium acetate trihydrate, pH 4.5) were harvested with the screen condition containing 18% glycerol. To obtain crystals of VP8*–HBGA complex, VP8* was co-crystallized with A-type trisaccharide or tetrasaccharide (purchased from Dextra labs), with a 1:52 or 1:46 excess molar ratio of ligand under similar condition as the unliganded P[14] VP8*.

Data collection and processing. Diffraction data for both unliganded and liganded VP8* crystals were collected at Baylor College of Medicine using Rigaku FR-E+ SuperBright rotating anode. These data were processed with DTREK[28] or IMOSFLM as implemented in the CCP4 suite[29]. Space group was confirmed using POINTLESS[30]. The unliganded and liganded VP8* structures in the $P2_1$ space group, with one molecule in the asymmetric unit, at ~1.5 Å resolution were determined. For initial phasing, the RRV VP8* structure (PDB ID 1KQR) was used as a search model for molecular replacement using Phaser[31]. Following automated model building and solvent addition using ARP/wARP[32], the structure was refined using PHENIX[33]. The oligosaccharide moieties of the HBGAs were generated using the SWEET2 package[34] of the Glycosciences.de server (http://www.glycosciences.de) and modelled into the electron density using COOT[35] and validated by computing simulated annealing omit maps using PHENIX[33]. The stereochemistry of the oligosaccharides including the allowed conformational angles was checked using the CARP[36] package in the Glycosciences.de server. Data collection and refinement statistics are given in Supplementary Table 2. Ligand interactions were analysed using COOT and LIGPLOT[37] with donor to acceptor distances between 2.6 Å and 3.2 Å for hydrogen-bonding interactions, and C-C distances between 3.4 Å and 4.5 Å for hydrophobic interactions. The structural alignments and calculations of r.m.s.d. were carried out using PyMOL (http://www.pymol.org/). Figures were prepared using Chimera[38].

Glycan array screening. The carbohydrate-binding specificity of HAL1166 VP8* was investigated on glycan array v4.2 comprised of 511 glycans (Consortium for Functional Glycomics, Protein-Glycan Interaction Core-H) (http://www.functionalglycomics.org). Recombinant GST-tagged VP8* at decreasing concentrations in binding buffer (20 mM Tris-HCl pH 7.4, 150 mM sodium chloride, 2 mM calcium chloride, 2 mM magnesium chloride, 0.05% Tween 20, 1% BSA) was applied to separate glycan arrays, and bound protein was detected using a fluorescent-labelled anti-GST monoclonal antibody. Summary of the glycan array results is given in Supplementary Table 1. Concentration dependent binding at 20 μg ml^{-1} and 2 μg ml^{-1} is shown in Supplementary Fig. 1a, b, where the glycans are ranked according to their relative binding strengths (Supplementary Fig. 1c) as described previously[39].

Inhibition and infectivity assays. Inhibition assays were performed on HT29 (human intestinal epithelial) cells following previously described protocols[40]. The monoclonal antibody (MAb) against blood group A antigen (BG-2) was purchased from Covance. The isotype control antibody (MG3-35) was purchased from Abcam. HAL1166 virus was grown as previously described[41]. Confluent HT29 cell monolayers were grown on 96-well plates. Increasing concentrations of antibodies were allowed to bind to the cells at 4 °C for 1 h before 400 fluorescent focus units (FFU) of virus were added per well. Recombinant VP8* protein competition was determined by treating the cells with 62.5 or 31 μg ml^{-1} VP8* at 4 °C for 30 min before virus inoculation as earlier. After allowing virus attachment to HT29 cells for 1 h on ice, the inoculum was removed, and the cells were washed with cold DMEM and incubated for 16 h at 37 °C in 95% (v/v) air with 5% (v/v) CO_2. Virus titres in methanol-fixed cell monolayers were determined by staining cells with rabbit anti-rotavirus antibody and Alexa 488-labelled donkey anti-rabbit secondary antibody (Invitrogen). Virus infectivity was expressed as the percentage of focus-forming units in control wells incubated with the same concentrations of bovine serum albumin. Data are given as the mean of three replicates, and the bar indicates the standard deviation. Data were analysed by Student's *t* test (two-tailed test).

For infectivity assays, SA11, HAL1166 or K8 viruses were serially diluted and incubated on confluent monolayers in 96-well plates of HT29, Caco-2 and CHO cells (parental cells or cells expressing type H HBGA, or type H and type A HBGA prepared as previously described[42]). Following attachment for 1 h at 37 °C, the cells were washed, incubated for 16 h and processed as described above for immunofluorescence.

Haemagglutination assay. Pooled human red blood cells were purchased from Immucor. The RBCs were packed via centrifugation at 500*g* for 10 min and 0.5% suspensions of each RBC type were prepared in 0.85% saline (pH 6.2). GST (negative control) or GST-tagged VP8* were serially diluted via doubling dilution in PBS (0.01 M sodium phosphate, 0.15 M NaCl, pH 5.5; passed through a 0.2-μm pore-size filter) on 96-well V-bottom plates (Nunc) in triplicate. The haemagglutination activity of the VP8 dimers was tested by mixing 50 μl of the prep (starting dilution of 10 nM) with an equal volume of the RBC suspensions. Recombinant norovirus virus-like particles (Norwalk virus, genogroup GI.1 and Houston virus, genogroup GII.4) were included as positive controls (5 μg ml^{-1} and 10 μg ml^{-1} starting dilutions, respectively) because they have well-characterized haemagglutination activity that is known to be mediated by interaction with histo-blood group antigens on the surface of the red blood cells[43]. The reaction was allowed to proceed for one hour at 4 °C before results were recorded. The titer recorded was the highest dilution of sample that prevented the complete sedimentation of red blood cells to the bottom of the well as compared to the negative controls. A commercial blood typing antibody specific for the B antigen was purchased from Immucor and tested at a starting dilution of 1:100 as another positive control.

28. Pflugrath, J. W. The finer things in X-ray diffraction data collection. *Acta Crystallogr. D* **55,** 1718–1725 (1999).
29. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
30. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
31. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
32. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP and automatic interpretation of protein electron density maps. *Methods Enzymol.* **374,** 229–244 (2003).
33. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
34. Bohne, A., Lang, E. & von der Lieth, C. W. SWEET - WWW-based rapid 3D construction of oligo- and polysaccharides. *Bioinformatics* **15,** 767–768 (1999).
35. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
36. Lütteke, T., Frank, M. & von der Lieth, C. W. Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.* **33,** D242–D246 (2005).
37. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.* **8,** 127–134 (1995).
38. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
39. Smith, D. F., Song, X. & Cummings, R. D. Use of glycan microarrays to explore specificity of glycan-binding proteins. *Methods Enzymol.* **480,** 417–444 (2010).
40. Haselhorst, T. *et al.* Sialic acid dependence in rotavirus host cell invasion. *Nature Chem. Biol.* **5,** 91–93 (2009).
41. Crawford, S. E. *et al.* Rotavirus viremia and extraintestinal viral infection in the neonatal rat model. *J. Virol.* **80,** 4820–4832 (2006).
42. Guillon, P. *et al.* Inhibition of the interaction between the SARS-CoV spike protein and its cellular receptor by anti-histo-blood group antibodies. *Glycobiology* **18,** 1085–1093 (2008).
43. Hutson, A. M., Atmar, R. L., Marcus, D. M. & Estes, M. K. Norwalk virus-like particle hemagglutination by binding to H histo-blood group antigens. *J. Virol.* **77,** 405–415 (2003).

LETTER

doi:10.1038/nature11016

Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia

Catherine C. Smith[1], Qi Wang[2], Chen-Shan Chin[3], Sara Salerno[1], Lauren E. Damon[1], Mark J. Levis[4], Alexander E. Perl[5], Kevin J. Travers[3], Susana Wang[3], Jeremy P. Hunt[6]†, Patrick P. Zarrinkar[6]†, Eric E. Schadt[3]†, Andrew Kasarskis[3]†, John Kuriyan[2] & Neil P. Shah[1,7]

Effective targeted cancer therapeutic development depends upon distinguishing disease-associated 'driver' mutations, which have causative roles in malignancy pathogenesis, from 'passenger' mutations, which are dispensable for cancer initiation and maintenance. Translational studies of clinically active targeted therapeutics can definitively discriminate driver from passenger lesions and provide valuable insights into human cancer biology. Activating internal tandem duplication (ITD) mutations in *FLT3* (*FLT3-ITD*) are detected in approximately 20% of acute myeloid leukaemia (AML) patients and are associated with a poor prognosis[1]. Abundant scientific[2] and clinical evidence[1,3], including the lack of convincing clinical activity of early FLT3 inhibitors[4,5], suggests that *FLT3-ITD* probably represents a passenger lesion. Here we report point mutations at three residues within the kinase domain of FLT3-ITD that confer substantial *in vitro* resistance to AC220 (quizartinib), an active investigational inhibitor of FLT3, KIT, PDGFRA, PDGFRB and RET[6,7]; evolution of AC220-resistant substitutions at two of these amino acid positions was observed in eight of eight *FLT3-ITD*-positive AML patients with acquired resistance to AC220. Our findings demonstrate that *FLT3-ITD* can represent a driver lesion and valid therapeutic target in human AML. AC220-resistant FLT3 kinase domain mutants represent high-value targets for future FLT3 inhibitor development efforts.

Perhaps the most compelling evidence so far that *FLT3-ITD* could represent a driver mutation in AML was the identification of a secondary *FLT3* kinase domain mutation that conferred moderate resistance to the multikinase inhibitor PKC412 in a single *FLT3-ITD*$^+$ patient who relapsed after an initial bone marrow response[8]. Although the broad-spectrum kinase inhibitor sorafenib has anecdotally achieved bone marrow remissions in *FLT3-ITD*$^+$ AML patients[9], whether its mechanism of action is mediated through inhibition of FLT3 or a distinct kinase is unclear. Indeed, two patients who relapsed after an initial response to sorafenib had no detectable *FLT3* kinase domain mutations at the time of resistance[10].

A recent interim analysis of 53 relapsed/refractory *FLT3-ITD*$^+$ AML patients evaluable for efficacy in a multinational phase II trial of AC220 monotherapy documented a composite complete remission (<5% bone marrow blasts) rate of 45% (frequently associated with incomplete recovery of peripheral blood counts)[7]. We sought to use the clinical activity of AC220 as a tool to define *FLT3-ITD* as a driver or passenger mutation in human AML. Using a previously validated *in vitro* saturation mutagenesis assay[11], we identified AC220 resistance-conferring mutations at four residues in the kinase domain of *FLT3-ITD* (Fig. 1a). Mutations at three of these amino acid positions conferred high degrees of *in vitro* AC220 resistance as demonstrated in proliferation (Fig. 1b) and cell-based biochemical assays (Fig. 1c). These residues consist of the 'gatekeeper' residue (F691) and two residues within the activation loop (D835, Y842). For unclear reasons, the E608K substitution did not confer substantial AC220 resistance and was not further characterized.

We next assessed the presence of drug-resistant kinase domain mutations in *FLT3-ITD* in eight paired pre-treatment and relapse samples obtained from *FLT3-ITD*$^+$ AML patients who initially achieved morphological reduction of bone marrow blasts to ≤5% with AC220 monotherapy, but subsequently relapsed despite continued AC220 treatment. In every case, subcloning and sequencing[12] of individual *FLT3-ITD* alleles revealed mutations at the time of relapse (Table 1) that were not detected pre-treatment (Supplementary Table 1). Mutations were confined to two of the three critical residues identified in our *in vitro* screen. The activation loop mutation D835Y was detected in three cases, D835V in two, and the gatekeeper mutation F691L was identified in three. Additionally, one novel activation loop mutation, D835F, was identified in a single patient. This mutation confers substantial *in vitro* resistance to AC220 (Supplementary Fig. 1) and cross-resistance to sorafenib (data not shown), and was probably

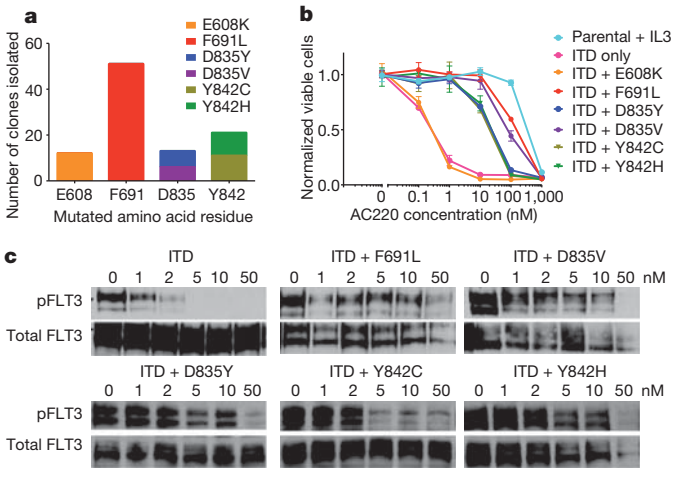


Figure 1 | Mutation screen of *FLT3-ITD* reveals secondary kinase domain mutations that cause varying degrees of resistance to AC220. a, Numbers of independent AC220-resistant Ba/F3 FLT3-ITD subpopulations with amino acid substitution at the indicated residue obtained from a saturation mutagenesis assay ($n = 97$ clones). **b**, Normalized cell viability of Ba/F3 populations stably expressing FLT3-ITD mutant isoforms after 48 h in various concentrations of AC220 (error bars represent standard deviations of triplicates from the same experiment). **c**, Western blot analysis using anti-phospho-FLT3 (pFLT3) or anti-FLT3 antibody performed on lysates from IL-3-independent Ba/F3 populations expressing the FLT3-ITD mutant isoforms indicated. Cells were exposed to AC220 at the indicated concentrations for 90 min.

[1]Division of Hematology/Oncology, University of California, San Francisco, California 94143, USA. [2]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [3]Pacific Biosciences, Menlo Park, California 94025, USA. [4]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, Maryland 21231, USA. [5]Abramson Cancer Center of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [6]Ambit Biosciences, San Diego, California 92121, USA. [7]Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, California 94115, USA. †Present addresses: KINOMEscan Division of DiscoveRx Corporation, San Diego, California 92121, USA (J.P.H.); Blueprint Medicines Corporation, Cambridge, Massachusetts 02142, USA (P.P.Z.); Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, New York 10029, USA (E.E.S. and A.K.).

Table 1 | Summary of FLT3 kinase domain mutations in patients relapsed on AC220

Subject number	Sex	Age (years)	Prior therapy	Karyotype at enrolment	Karyotype at relapse	Blasts in relapse sample (%)	New mutation at relapse	ITD$^+$ clones with mutation	Weeks on study
1009-003	F	75	7+3	45~54,XX,+3,+6,+7, +8,+13,+14,+21, +22[cp15]/46,XX[5]	52,XX,+3,+6,+7,+8, +10,+12,+13[cp7]/46, XX[14]	90	D835F	6/15	12
1011-006	M	70	7+3, low-dose cytarabine	Normal	ND	10	D835Y	4/15	8
1011-007	F	56	7+3, HAM	Normal	46,XX,del(11) (p?13p?15) [12]/46,XX[9]	80	F691L D835V	4/24 5/24	11
1005-004	F	60	Cytarabine and mitoxantrone	Normal	Normal	92	F691L	9/22	19
1005-006	M	43	7+3, MEC, allogeneic stem cell transplant	6,XY,t(1;15)(p22;q15)	ND	59	D835Y	8/17	6
1005-007	F	59	7+3, HDAC	Normal	ND	39	D835V	9/21	23
1005-009	M	68	Cytarabine and mitoxantrone	Normal	ND	58	D835Y	8/14	19
1005-010	M	52	7+3, HDAC, mitoxantrone and etoposide	46,XY,t(4;12)(q26;p11.2), t(8;14)(q13;q11.2)	ND	22	F691L	6/18	20

All patients achieved morphological bone marrow blasts of ≤5% at best response. 7+3, low-dose cytarabine for 7 days plus 3 days anthracycline; HAM, high-dose cytarabine plus mitoxantrone; HDAC, high-dose cytarabine; MEC, mitoxantrone, etoposide, cytarabine. ND, not done.

not detected in our saturation mutagenesis screen because its creation requires a two-nucleotide substitution. One patient (1011-007) seemed to have evolved polyclonal resistance, with both F691L and D835V mutations detected on separate *FLT3-ITD* sequences. Collectively, these findings suggest that clinical response and relapse in each of these eight patients is mechanistically mediated through modulation of FLT3-ITD kinase activity.

To assess more precisely for resistance-conferring mutations at relapse, we used a recently described single molecule real-time (SMRT; Pacific Biosciences) sequencing platform, which can provide sequencing reads of sufficient length to enable focused interrogation of *FLT3-ITD* alleles (Supplementary Fig. 2)[13]. With this assay, hundreds of reads (range 19–930) spanning the ITD region and kinase domain with an average read length of greater than 1 kilobase (kb) (Supplementary Table 2) were reliably obtained from individual patient samples. Attention was focused on the amino acid codons identified in the *in vitro* screen for AC220 resistance-conferring mutations. SMRT sequencing confirmed the presence of resistance-conferring kinase domain mutations in *FLT3-ITD* at relapse in all eight patient samples (Table 2). Consistent with the results obtained by subcloning and sequencing, mutations at E608 and Y842 were not detected. The frequency of individual alternative codon substitutions within *FLT3-ITD* ranged from as low as 2.7% (D835F in patient 1005-006) to 50.6% (D835Y in patient 1005-009). The presence of polyclonal resistance was confirmed in patient 1011-007, and noted in an additional three cases: 1009-003, 1005-006 and 1005-007 (Table 2 and Supplementary Fig. 3). In general, mutations were detected on distinct molecules, although in the case of 1011-007, a subset of *FLT3-ITD* molecules with F691L also harboured a D835V mutation (5/21 observations; 23.8% of *FLT3-ITD(F691L)* alleles; data not shown). Analysis of three normal control samples revealed base substitutions at these residues at a very low frequency (Table 2 and Supplementary Table 3). The evolution of polyclonal resistance due to secondary kinase domain mutations in *FLT3-ITD* in four of eight relapsed patients is supportive of a central dependence upon FLT3-ITD signalling in the leukaemic clone of a subset of AML patients, and indicative of profound selective pressure exerted upon this clone by AC220. Additionally, these findings reveal the genetic complexity of drug-resistant disease that may evolve in cancer patients on clinically active therapy.

All mutations identified at relapse were detected at a frequency significantly higher than that observed in a normal control, and although relapse occurred relatively rapidly in some patients, mutations were not convincingly detectable before treatment. The aggregate frequency of all mutations at relapse in individual patients ranged from approximately 20–50% in all cases, which is consistent with leukaemic blasts homozygous for *FLT3-ITD* and containing one drug-resistant allele per cell, although the presence of a heterogeneous blast population with only a subset of drug-resistant *FLT3-ITD*$^+$ cells expressing kinase domain mutations cannot be excluded.

The five substitutions that conferred a high degree of resistance to AC220 *in vitro* were cross-resistant to sorafenib in cell-based growth and biochemical assays (Supplementary Fig. 4). The degree of relative resistance to sorafenib associated with these mutants was generally similar to that observed with AC220 (Supplementary Table 4).

Table 2 | Third-generation sequencing identifies polyclonal *FLT3* kinase domain mutations

				Pre-treatment		Relapse		Normal control no. 1	
Subject number	Mutation	Native codon	Alternative codon	Observed alternative codon frequency in ITD$^+$ sequences (%)	Total number of ITD$^+$ sequences sampled	Observed alternative codon frequency in ITD$^+$ sequences (%)	Total number of ITD$^+$ sequences sampled	Observed alternative codon frequency (%)	Total number of sequences sampled
1009-003	D835Y	GAT	TAT	0.21	482	8.4	332	0.00	768
	D835V	GAT	GTT	0.00	482	3.3	332	0.13	768
	D835F	GAT	TTT	0.00	482	10.2	332	0.00	768
1011-006	D835Y	GAT	TAT	0.00	196	41.0	402	0.00	768
1011-007	F691L	TTT	TTG	0.18	561	6.2	341	0.22	450
	D835Y	GAT	TAT	0.00	930	3.0	436	0.00	768
	D835V	GAT	GTT	0.43	930	29.6	436	0.13	768
1005-004	F691L	TTT	TTG	0.00	496	29.6	513	0.22	450
1005-006	D835Y	GAT	TAT	0.00	171	39.5	261	0.00	768
	D835F	GAT	TTT	0.00	171	2.7	261	0.00	768
1005-007	D835Y	GAT	TAT	0.00	57	4.0	378	0.00	768
	D835V	GAT	GTT	0.00	57	47.4	378	0.13	768
1005-009	D835Y	GAT	TAT	0.00	19	50.6	445	0.00	768
1005-010	F691L	TTT	TTG	0.00	387	25.3	150	0.22	450

All P values $<1 \times 10^{-5}$ for alternative codon frequencies at relapse compared to a representative normal control sample (no. 1 refers to one of three normal control samples analysed).

To understand the structural effects of AC220-resistance conferring mutations, we modelled the binding of AC220 to FLT3 (Fig. 2a). The crystal structure of the FLT3 kinase domain has been determined previously in an inactive conformation[14] that resembles the inactive conformations of ABL[15], KIT[16] and insulin receptor tyrosine kinase[17], with the activation loop folded back onto the ATP-binding cleft (loop-in conformation), thereby preventing substrate loading. The Asp-Phe-Gly (DFG) motif at the amino-terminal end of the activation loop adopts the DFG-out conformation, in which the Asp side chain, which normally coordinates a magnesium ion, is removed from the active site. The activation of FLT3 would require flipping of the DFG motif and reorganization of the activation loop, as observed in ABL[18] and insulin receptor kinase[19]. Previously published binding data suggest that AC220 is a type II kinase inhibitor that preferentially binds to the inactive, DFG-out kinase conformation[20]. Our molecular docking analysis supports a model whereby AC220 interacts favourably with the DFG-out, inactive conformation. In the docked AC220–FLT3 model, the AC220 amide group is 2.6 Å from the carbonyl group of C694, consistent with the formation of a hydrogen bond. The phenol ring of AC220 forms a perpendicular aromatic–aromatic interaction[21] with F830 in the DFG motif (Fig. 2b). This interaction would not be possible in the DFG-in, active kinase conformation. The gatekeeper residue F691 forms a π–π stacking contact with the benzo-imidazol-thiazol moiety of AC220, which may further stabilize the complex. Substitutions at F691 with non-aromatic residues such as leucine may not compensate for the π–π stacking interaction.

Residues D835 and Y842 stabilize the inactive conformation of the activation loop by forming hydrogen bonds with the main-chain amide group of S838 and the side chain of D811, respectively (Fig. 2c). Thus, replacement of either residue might destabilize this particular inactive conformation of the activation loop, which would then be expected to hinder the binding of AC220. In further support of this model, a binding study of AC220 and sorafenib, a crystallographically proven type II inhibitor[22], revealed that the binding affinity of both inhibitors to the FLT3 D835V mutant is substantially reduced compared to native FLT3, both in the presence and absence of the juxtamembrane domain containing the ITD (Supplementary Fig. 5 and Supplementary Table 5).

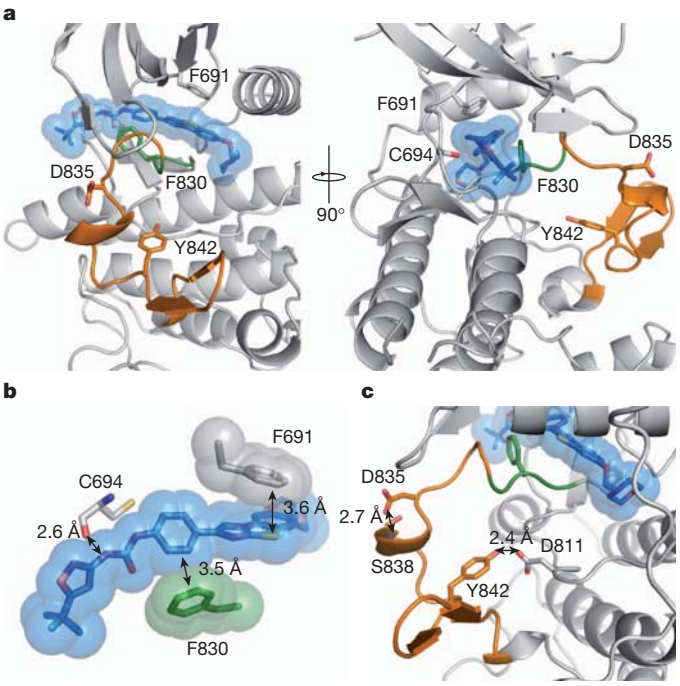


Figure 2 | Modelling of FLT3–AC220 interactions. **a**, Docking model of the AC220-bound FLT3 kinase domain: AC220 (blue); activation loop (orange) and DFG motif (green); amino acid residues that confer AC220 resistance when mutated (F691, D835 and Y842) and that interact with AC220 (F691, C694 and F830) in sticks. **b**, Surface and stick presentation of AC220 and AC220-interacting residues on FLT3: the carbonyl oxygen of C694 interacts with one of the AC220 amide groups; F691 forms a π–π stacking interaction with AC220; F830 interacts with AC220 through a perpendicular aromatic–aromatic interaction. **c**, Structure of the activation loop: residues D835,Y842 and interacting residues on FLT3 depicted in sticks.

Other potential explanations for the mechanism of resistance conferred by AC220-resistant mutants include increased kinase activity and differential activation of downstream effectors. Western blot analysis of cells expressing AC220-resistant FLT3-ITD-mutant isoforms revealed increased FLT3 autophosphorylation of D835 mutant isoforms, but no discernable difference in phosphorylation of downstream targets, including the direct FLT3-ITD target STAT5A/B[23], and no difference in cellular proliferation (Supplementary Fig. 6). Overall, these data support a primarily structural mechanism for AC220 resistance mediated by mutations at F691, D835 and Y842, although further studies are necessary for definitive confirmation. We speculate that the ability to retain inhibitory activity against activation loop substitutions at D835 and Y842 will require a type I FLT3 kinase inhibitor capable of effectively binding to the active, DFG-in conformation of the kinase.

Substitutions at gatekeeper residues such as FLT3-ITD(F691) are well-documented causes of resistance to kinase inhibitors[12,24]. Analogues of the FLT3-ITD(D835V) activation loop mutation have proven problematic for a number of kinase inhibitors: KIT(D816V), an activating mutation that is commonly detected in systemic mastocytosis, confers a high degree of resistance to imatinib and other KIT inhibitors[25]. Our data, although derived from a small cohort of patients that will need to be validated in larger studies, suggest that substitutions at F691 and D835 in FLT3-ITD will pose substantial barriers to disease control in AML patients treated with either AC220 or sorafenib, and therefore represent high-value targets for novel FLT3 inhibitor development efforts.

Compelling data suggest that activating *FLT3* mutations are acquired relatively late during leukaemogenesis in a pre-established clone[1,3], and alone are insufficient to cause acute leukaemia in pre-clinical models[2]. Recent evidence suggests that the molecular heterogeneity of individual leukaemias can be substantial, and can occur in both branching and linear fashions early during leukaemogenesis, including at the leukaemia-initiating or 'leukaemic stem' cell level[26]. Collectively, our data are consistent with acquisition of *FLT3-ITD* and drug-resistant *FLT3* kinase domain mutations in a leukaemia-initiating cell population, although formal transplantation studies in mice are needed to address this question definitively. Our findings validate FLT3-ITD as a therapeutic target in human AML, and suggest that FLT3-ITD is capable of conferring a state of 'oncogene addiction', whereby cellular survival pathways associated with normal or precancerous cells can become hijacked, leading to a state of reliance upon key signalling molecules that can be exploited therapeutically. This work supports the exploration of therapeutic strategies targeting select activating mutations in other signalling molecules that are believed to be acquired relatively late in disease evolution, such as JAK2 (ref. 27) or RAS[3], with agents capable of achieving clinically meaningful target inhibition. Further studies will be required to identify mechanisms of drug resistance that may circumvent reliance on activated FLT3 by activation of downstream or parallel pathways, as has been described with other kinase inhibitors[28]. To that end, translational studies using detailed molecular analyses of primary samples obtained from AML patients treated with clinically effective targeted therapeutics promise to further inform mechanisms of drug resistance, strategies for future drug development, and models of disease evolution.

METHODS SUMMARY

MSCVpuroFLT3-ITD plasmid DNA was mutagenized and used to generated AC220-resistant Ba/F3 clones as previously described[11]. The *FLT3* kinase domain

was sequenced from PCR-amplified genomic DNA isolated from AC220-resistant clones. Identified drug-resistant mutations were re-engineered into MSCVpuroFLT3-ITD using site-directed mutagenesis and Ba/F3 cell lines were created as detailed in Methods. Cell viability in the presence and absence of drug was assessed using trypan blue exclusion. FLT3 phosphorylation status was determined by western blot analysis of whole cell lysates prepared after 90 min of drug exposure from Ba/F3 cells stably expressing FLT3 mutant isoforms and from transfected 293T cells in the absence of drug. The *FLT3* kinase domain was PCR amplified from cDNA derived from blood or bone marrow samples from patients enrolled on the exploratory portion of the phase II trial of AC220 in AML (http://clinicaltrials.gov/ct2/show/NCT00989261; identifier NCT00989261) and from normal control bone marrow or mobilized peripheral blood stem cells. PCR products were cloned into *Escherichia coli* and individual clones were sequenced using Sanger sequencing. Alternatively, SMRTBell libraries[13] were prepared as per the manufacturer's instructions and sequenced on a Pacific Biosciences RS instrument. For details of the computational sequencing analysis, please see Methods. Molecular docking of FLT3-ITD to AC220 was performed using Autodock 4.2 package. Inhibitor binding constants were measured using an active-site-dependent competition binding assay as previously descibed[20]. For further details of methods, please see Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 August 2011; accepted 5 March 2012.
Published online 15 April 2012.

1. Thiede, C. *et al.* Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood* **99,** 4326–4335 (2002).
2. Lee, B. H. *et al.* FLT3 internal tandem duplication mutations induce myeloproliferative or lymphoid disease in a transgenic mouse model. *Oncogene* **24,** 7882–7892 (2005).
3. Shih, L. Y. *et al.* Acquisition of *FLT3* or N-*ras* mutations is frequently associated with progression of myelodysplastic syndrome to acute myeloid leukemia. *Leukemia* **18,** 466–475 (2004).
4. Knapper, S. *et al.* A phase 2 trial of the FLT3 inhibitor lestaurtinib (CEP701) as first-line treatment for older patients with acute myeloid leukemia not considered fit for intensive chemotherapy. *Blood* **108,** 3262–3270 (2006).
5. Fischer, T. *et al.* Phase IIB trial of oral Midostaurin (PKC412), the FMS-like tyrosine kinase 3 receptor (FLT3) and multi-targeted kinase inhibitor, in patients with acute myeloid leukemia and high-risk myelodysplastic syndrome with either wild-type or mutated FLT3. *J. Clin. Oncol.* **28,** 4339–4345 (2010).
6. Zarrinkar, P. P. *et al.* AC220 is a uniquely potent and selective inhibitor of FLT3 for the treatment of acute myeloid leukemia (AML). *Blood* **114,** 2984–2992 (2009).
7. Cortes, J. *et al.* in *16th Congress of the European Hematology Association* (Haematologica, 2011).
8. Heidel, F. *et al.* Clinical resistance to the kinase inhibitor PKC412 in acute myeloid leukemia by mutation of Asn-676 in the FLT3 tyrosine kinase domain. *Blood* **107,** 293–300 (2006).
9. Metzelder, S. *et al.* Compassionate use of sorafenib in FLT3-ITD-positive acute myeloid leukemia: sustained regression before and after allogeneic stem cell transplantation. *Blood* **113,** 6567–6571 (2009).
10. Scholl, S. *et al.* Secondary resistance to sorafenib in two patients with acute myeloid leukemia (AML) harboring FLT3-ITD mutations. *Ann. Hematol.* **90,** 473–475 (2011).
11. Azam, M., Latek, R. R. & Daley, G. Q. Mechanisms of autoinhibition and STI-571/imatinib resistance revealed by mutagenesis of *BCR-ABL*. *Cell* **112,** 831–843 (2003).
12. Shah, N. P. *et al.* Multiple *BCR-ABL* kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell* **2,** 117–125 (2002).
13. Travers, K. J., Chin, C. S., Rank, D. R., Eid, J. S. & Turner, S. W. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.* **38,** e159 (2010).
14. Griffith, J. *et al.* The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. *Mol. Cell* **13,** 169–178 (2004).
15. Levinson, N. M. *et al.* A Src-like inactive conformation in the Abl tyrosine kinase domain. *PLoS Biol.* **4,** e144 (2006).
16. Mol, C. D. *et al.* Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem.* **279,** 31655–31663 (2004).
17. Hubbard, S. R., Wei, L., Ellis, L. & Hendrickson, W. A. Crystal structure of the tyrosine kinase domain of the human insulin receptor. *Nature* **372,** 746–754 (1994).
18. Mol, C. D. *et al.* Structure of a c-Kit product complex reveals the basis for kinase transactivation. *J. Biol. Chem.* **278,** 31461–31464 (2003).
19. Hubbard, S. R. Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J.* **16,** 5572–5581 (1997).
20. Wodicka, L. M. *et al.* Activation state-dependent binding of small molecule kinase inhibitors: structural insights from biochemistry. *Chem. Biol.* **17,** 1241–1249 (2010).
21. Burley, S. K. & Petsko, G. A. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **229,** 23–28 (1985).
22. Wan, P. T. *et al.* Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116,** 855–867 (2004).
23. Choudhary, C. *et al.* Activation mechanisms of STAT5 by oncogenic Flt3-ITD. *Blood* **110,** 370–374 (2007).
24. Pao, W. *et al.* Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Med.* **2,** e73 (2005).
25. Barbie, D. A. & Deangelo, D. J. Systemic mastocytosis: current classification and novel therapeutic options. *Clin. Adv. Hematol. Oncol.* **4,** 768–775 (2006).
26. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469,** 356–361 (2011).
27. Kralovics, R. *et al.* Acquisition of the V617F mutation of *JAK2* is a late genetic event in a subset of patients with myeloproliferative disorders. *Blood* **108,** 1377–1380 (2006).
28. Nazarian, R. *et al.* Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* **468,** 973–977 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Lin for technical assistance. This work was funded by grants from the Leukemia and Lymphoma Society (to C.C.S. and N.P.S.), the Doris Duke Charitable Foundation (to N.P.S.), NCI Leukemia SPORE P50 CA100632-06 (to M.J.L.), NCI R01 CA12886 (to M.J.L.) and the NIH T-32 Molecular Mechanisms of Cancer (to C.C.S.). C.C.S. would like to acknowledge the EHA/ASH Translational Research Training Institute. N.P.S. would like to thank Art and Alison Kern and the Edward S. Ageno family for their support.

Author Contributions C.C.S., Q.W., C.-S.C., K.J.T., A.K., E.E.S. and J.K. designed experiments, performed research, analysed data and wrote the manuscript. N.P.S. designed experiments, analysed data and wrote the manuscript. L.E.D., S.W., J.P.H. and S.S. performed experiments and reviewed the manuscript. P.P.Z. was involved in study design and reviewed the manuscript. A.E.P. and M.J.L. provided reagents, performed research and reviewed the manuscript.

Author Information SMRT sequencing data is deposited online at http://www.ncbi.nlm.nih.gov/sra under accession number SRA050226.1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to N.S. (nshah@medicine.ucsf.edu).

METHODS

DNA constructs, mutagenesis and resistance screen. *FLT3-ITD* cDNA cloned from the MV4;11 cell line (ITD: residues 591–601) into the HpaI site of the pMSCVpuro retroviral vector (Clontech) was a gift from Ambit Biosciences and was used as a template for mutagenesis. We used a modified strategy for random mutagenesis previously described[11]. Briefly, 1 μg of MSCVpuroFLT3-ITD was used to transform the DNA-repair-deficient *Escherichia coli* strain XL-1 Red (Stratagene) and plated on 20 ampicillin-agar bacterial plates. After incubation for 36 h, colonies were collected by scraping, and plasmid DNA was purified by using a plasmid MAXI kit (Qiagen). Subsequently, mutagenized *FLT3-ITD* plasmid stock and Ecopack packaging plasmid were cotransfected into 293T cells grown in DMEM (Invitrogen) containing 10% fetal calf serum (FCS; Omega Scientific) using Lipofectamine 2000 (Invitrogen) as per the manufacturer's protocol. Viral supernatants were collected at 48 h, purified using a 0.44 μm vacuum filter, and used to infect Ba/F3 cells at a 1:100 to 1:300 dilution of viral supernatant to fresh RPMI 1640 (Invitrogen) supplemented with 10% FCS. Alternatively, viral supernatant was aliquoted and frozen. Thawed supernatant was used to infect Ba/F3 cells at a 1:50 dilution. Viral supernatant was diluted with the goal of minimizing multiplicity of infection. For infection, $1–2 \times 10^6$ Ba/F3 cells were resuspended in 3 ml of the diluted viral stock supplemented with recombinant mouse IL-3 (Invitrogen), and 4 μg ml^{-1} polybrene, plated in each well of a 12-well tissue culture dish and centrifuged at 1,500*g* in a Beckman Coulter Allegra 6KR centrifuge with a microplate carrier for 90 min at 34 °C. Centrifuged cells were subsequently transferred to a 37 °C incubator overnight. Infected Ba/F3 cells were washed twice with media to remove IL-3 and plated in 3 ml of RPMI medium 1640 at 5×10^5 cells per well of a 6-well dish supplemented with 20% FCS and 1.2% Bacto-agar with 20 nM AC220 (a gift from Ambit Biosciences). After 10–21 days, visible colonies were plucked from agar and expanded in the presence of drug (20 nM AC220).

Sequencing and alignments. Expanded colonies were harvested 7–14 days after isolation from agar, and whole genomic DNA was isolated using the QIAamp kit (Qiagen). The *FLT3* kinase domain was amplified by PCR from whole genomic DNA by using TopTaq DNA polymerase (Qiagen). The primers TK1F (5′- TG CTGTGCATACAATTCCCTTGGC-3′) and TK2R (5′- TCTCTGCTGAAAGGT CGCCTGTTT-3′) were used for kinase domain amplification and subsequent bidirectional sequencing was performed using these primers in addition to TK1R (5′- AGTCCTCCTCTTCTTCCAGCCTTT-3′) and TK2F (5′- GAGA GGCACTCATGTCAGAACTCA-3′). Alignments to the native *FLT3-ITD* sequence were performed using Sequencher software (Gene Codes Corporation).

Generation of mutants. Mutants isolated in the screen were engineered into pMSCVpuroFLT3-ITD by using the QuikChange mutagenesis kit (Stratagene). In all cases, individual point mutants were confirmed by sequence analysis.

Cell viability and proliferation assays. Stable Ba/F3 lines were generated by using retroviral spinfection with the appropriate mutated plasmid as outlined above, with the exception of the exclusion of polybrene. At 48 h post-infection, puromycin was added to infected cells at a concentration of 4 μg ml^{-1}. Cells were selected in the presence of puromycin for 7–10 days and subsequently IL-3 was washed twice from the cells with media and cells were selected in RPMI medium 1640 plus 10% FCS in the absence of IL-3. Exponentially growing Ba/F3 cells (5×10^4) were plated in each well of a 24-well dish with 1 ml of RPMI 1640 plus 10% FCS containing the appropriate concentration of drug as indicated in triplicate. Cells were allowed to expand for 2 days and were counted by using a Vi-cell XR automated cell viability analyser (Beckman Coulter). The mean number of viable cells at varying concentrations of drug was normalized to the median number of viable cells in the no-drug sample for each mutant. Error bars represent the standard deviation. Numerical IC_{50} values were generated using nonlinear best-fit regression analysis using Prism 5 software (GraphPad).

For proliferation assays, on day 0, parental Ba/F3 cells and Ba/F3 cells stably expressing FLT3-ITD mutant isoforms were plated in triplicate with 1 mL of RPMI 1640 plus 10% FCS at a density of 5×10^4 cells per well in each well of a 24-well dish. Cells were allowed to expand and were counted by using a Vi-cell XR automated cell viability analyser (Beckman Coulter) on days 2, 3, 4, 5, 6 and 7. To maintain exponential growth of cells, 0.5 ml of cells from each well were used for counting on each day and 0.25 ml of cells from the remaining volume were transferred to a new well with 0.75 ml of fresh RPMI plus 10% FCS (including 2 ng ml^{-1} of IL-3 for parental Ba/F3 cells). Extrapolated cell counts were calculated from the measured count on each day using the appropriate dilution factor (1× on day 2, 4× on day 3, 16× on day 4, and so on). The number of viable cells at each time point was normalized to the starting number of cells for each cell line on day 0 and the mean normalized cell count on each day was calculated. Error bars represent the standard deviation.

Immunoblotting. Exponentially growing Ba/F3 cells stably expressing each mutation along with a native FLT3-ITD control were plated in RPMI medium 1640 plus 10% FCS supplemented with kinase inhibitor at the indicated concentration. After a 90-min incubation, the cells were washed in phosphate buffered saline (PBS) and lysed in Cell Extraction Buffer (Invitrogen) supplemented with protease and phosphatase inhibitors. The lysate was clarified by centrifugation and quantified by BCA assay (Thermo Scientific). Protein was subjected to sodium dodecylsulphate polyacrylamide electrophoresis and transferred to nitrocellulose membranes. Immunoblotting was performed using anti-phospho-FLT3, anti-phospho-STAT5, anti-STAT5, anti-phospho-ERK, anti-ERK, anti-phospho-S6, anti-S6, anti-GAPDH (Cell Signaling) and anti-FLT3 S18 antibody (Santa Cruz Biotechnology). Alternatively, 293T cells were plated in 6-cm plates and transfected with MSCVpuroFLT3-ITD plasmid containing FLT3 mutations of interest using Lipofectamine 2000 (Invitrogen) as per the manufacturer's protocol. After 48 h, cells were washed with PBS, collected, lysed and subjected to western blot analysis as described above.

Competition binding assays. Inhibitor binding constants were measured by using active site-dependent competition binding assays essentially as previously described[20]. In brief, FLT3 protein isoforms were labelled with a chimaeric double-stranded DNA tag containing the NFkB binding site (50-GGGAATTCCC-30) fused to an amplicon for qPCR readout, which was added directly to the expression extracts. Binding reactions were assembled by combining DNA-tagged kinase extract, affinity beads loaded with a kinase inhibitor probe molecule, and test compound in 13 binding buffer (PBS, 0.05% Tween 20, 10 mM DTT, 0.1% BSA, 2 mg ml^{-1} sonicated salmon sperm DNA). Extracts were used directly in binding assays without any enzyme purification steps at a ≥10,000-fold overall stock dilution (final DNA-tagged enzyme concentration <0.1 nM). Assays were incubated for 1 h at room temperature (23 °C), which was sufficient to establish equilibrium. Subsequent washing, elution, and qPCR readout steps were as described[20].

Patients and FLT3 kinase domain sequencing analysis. Eight cases of acquired resistance to AC220 were analysed. Patients were enrolled on the exploratory cohort of the phase II clinical trial of AC220 in relapsed or refractory AML at the University of California, San Francisco (UCSF), University of Pennsylvania or Johns Hopkins University. Details of the clinical trials and results are reported elsewhere[7]. All patients were *FLT3-ITD*-positive at enrolment. Samples were collected pre-treatment and at the time of disease progression. Only patients who had achieved morphological clearance of bone marrow blasts to ≤5% at best response and subsequently relapsed with an increase in peripheral blood or bone marrow blasts are included in this analysis. The patients in this analysis included all the patients meeting the above criteria at the three participating institutions. All patients gave informed consent according to the Declaration of Helsinki to participate both in the clinical trials and for collection of samples. All research involving human subjects was approved by the relevant Institutional Review Board at each individual participating institution (UCSF, University of Pennsylvania or Johns Hopkins).

For sequencing, frozen Ficoll-purified mononuclear cells obtained from blood or bone marrow were lysed in Trizol (Invitrogen) and RNA was isolated according to the manufacturer's protocol. cDNA was synthesized using Superscript II (Invitrogen) as per the manufacturer's protocol. The *FLT3* kinase domain and adjacent juxtamembrane domain were PCR amplified from cDNA using primers TK1F and TK2R as above. PCR products were cloned using TOPO TA cloning (Invitrogen) and transformed into competent *E. coli*. Individual colonies were plucked, expanded in liquid culture overnight and plasmid DNA for sequencing was isolated using the QIAprep Spin Miniprep kit (Qiagen). Each colony was considered representative of a single mRNA. To minimize contamination from PCR artefact, we sequenced at least 10 and up to 24 *FLT3-ITD*-containing clones from each sample and required that mutations be found in >15% of clones. The primers TK1F, TK1R, TK2F and TK2R were used for bidirectional sequencing as above. Alignments with native *FLT3* sequence were performed using Sequencher software (Gene Codes Corporation).

Sample preparation and SMRT sequencing. PCR product containing the *FLT3* kinase domain was generated from patient cDNA as described above using high fidelity DNA polymerase. We prepared PCR products for Pacific Biosciences sequencing using standard commercial kits and reagents (http://www.pacificbiosciences.com/products/consumables/reagents) following the manufacturer's instructions. PCR products input amounts ranged from 0.3–3 μg, and we prepared SMRTBell libraries[13] on the full PCR products without any fragmentation. We sequenced all samples on a Pacific Biosciences RS instrument and recorded sequence for 75 min.

Computational analysis of FLT3 mutations. We obtained samples from three healthy individuals with no cancer history (two bone marrow, one mobilized peripheral blood stem cells), isolated RNA, made cDNA, amplified the *FLT3* kinase domain, and sequenced following a protocol identical to that used on the AML samples. We used the sequence from Normal Control no. 1 as a control for all process steps between sample acquisition and sequencing. Data from the remaining two normal controls were compared to the Normal Control no. 1

and revealed no significant differences (Supplementary Table 3). We use the circular consensus sequencing (CCS) mode to obtain high accuracy reads for the ~1.4 kb amplicon with PacBio RS. The CCS mode generates reads by combining multiple independent single-pass sequencing reads for individual molecules to correct raw errors and generate a better accuracy consensus (see Supplementary Fig. 2). We report only the CCS reads where the same molecule is sequenced at three or more times, that is, raw read length >4.2 kb for the 1.4 kb amplicon. With the CCS reads, we obtained the sequence of the ~1.4 kb amplicon containing the FLT3 juxtamembrane and kinase domains with up to about 98% to 99% accuracy (see alignment identity for ITD^- samples in Supplementary Table 2). For each CCS read, we used tandem repeats finder (TRF)[29] to identify the ITD sequence. To determine unambiguously whether a read was ITD^- or ITD^+ consistently, we used only the CCS reads that included at least the region from the 50-bp 5′-end upstream to the 50-bp 3′-end downstream sequence of the ITD region in the analysis. This allowed us to determine the number of sequences containing the ITD more accurately despite a small percentage of insertion and deletion errors in the CCS reads. Two distinct peaks allowed us to identify ITD^- versus ITD^+ CCS reads unambiguously. We found that each sample had only one major ITD as expected, although in some cases the majority ITD differed at relapse compared to pre-treatment. We then passed the ITD^+ population of the CCS reads to the next stage for codon mutation analysis. A list of the number of total CCS reads identified is listed in Supplementary Table 2. We identified ~200–1,300 CCS reads spanning the whole region between the ITD region and the furthest codon of interest (Y842) for codon analysis per sample.

For codon mutation analysis, we restricted our analysis to the 608, 691, 835 and 842 codons from reference sequence NM_004119 (*Homo sapiens FLT3* mRNA) and then took the frequency of sequences obtained for each of these codons in the PCR amplicon of healthy Normal Control no. 1 and compared that to the frequency of sequences in each AML patient sample. A local quality filter that required exact matching of the codons before and after the codon of interest was used for filtering out low quality codon calls that might be due to sequencing errors. We used the observed frequencies from the control sample for calculating the significance of the observed mutation in the AML patient samples. The *P* value was calculated by comparing the numbers of native codons observed and the alternative codon between the control sample and the AML patient sample with Fisher's exact test on the contingency table[30,31]. Owing to the potential statistical bias that could arise if the number of observed mutations was small in some cases, or if sequencing error frequencies differed between mutant and reference codon sequences, we only report the mutations using a conservative significance threshold of $P < 1 \times 10^{-5}$. We used a simulation to determine the sensitivity of this analysis to detect a true mutation at a given codon position. Sensitivity in this analysis was determined as a function of three parameters: the number of errors observed in the control sequence, the total number of times that position is sequenced in the control, and the number of times that position is sequenced in the patient sample. For the simulation, we conservatively assumed that all alternative codons seen in the control are actual errors. With this simulation, we estimated that this analysis allows us to detect variants in the patient sample above 3% with high confidence if we get more than 300 observations of the codon of interest. To refine further our search for mutations underlying relapse in these patients, we considered only those mutations that were in *cis* to an ITD, as defined on being on the same single DNA molecule sequence read. These mutations at both baseline and relapse are listed in Table 2.

Molecular docking. Molecular docking was performed using Autodock 4.2 package[32]. The FLT-ITD structure (residue 587–947) was prepared from the Protein Data Bank accession 1RJB[14]. All bound waters were removed from the protein. The structure was then added for hydrogens, and partial atomic charges were assigned using AutoDockTools (ADT)[32]. Residues K644, F830, F691 and E661 were selected as flexible residues. To define the flexible residues, we first analysed the crystal structure of imatinib-bound inactive KIT kinase domain. The structure of FLT3 is quite similar to that of KIT, so this comparison helps us identify potential ligand interacting residues in FLT3. In the KIT structure, residue L595, K623, E640, L644, T670, Y672, L799 and F811 are close to the ligand. We thus define the corresponding residues in the FLT3 kinase domain (Y693, F830, F691, K644, E661, L818, L616, M665, V624) as flexible. Our docking studies revealed that the conformations of Y693, L818, L616, M665 and V624 in the docking solutions are largely identical to their conformations in the crystal structure, and so we did not consider these five residues to be flexible in the final calculations.

The coordinates of AC220 were generated using the Dundee PROGRD2 server[33], and its initial conformation was energy minimized by the GROMACS force field. The Gasteiger charges were then assigned to the ligand using ADT. Seven torsion bonds were defined as rotable during the docking procedure. The ligand was put into the kinase ATP-binding pocket and manually aligned to avoid atom clashes. A three-dimensional grid box (dimensions: $60 \times 30 \times 60$ unit in number of grid points, grid spacing: 0.375 Å) centred at the ligand defining the search space was then created by AutoGrid4.2 (ref. 32). Two hundred runs of Larmarckian Genetic Algorithm were performed to optimize the ligand–protein interactions. The solutions were clustered according to the root mean standard deviation values, and ranked by the binding free energy. Two general poses are observed. The top-ranked pose has an average energy of $-10.32\ \text{kcal mol}^{-1}$ (the lowest energy for this pose is $-10.93\ \text{kcal mol}^{-1}$). The second-ranked pose, which is flipped by 180° with respect to the top-ranked position, has 133 solutions (63%) with an average energy of $-5.82\ \text{kcal mol}^{-1}$ (the lowest energy for this pose is $-6.98\ \text{kcal mol}^{-1}$). Given the gap in the calculated energy, we only picked the lowest one for the purely illustrative purposes of this analysis.

29. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
30. Yates, F. Tests of significance for 2 × 2 contingency tables. *J. R. Stat. Soc. A* **147,** 426–463 (1984).
31. Barnard, G. A. Must clinical trials be large? The interpretation of *p*-values and the combination of test results. *Stat. Med.* **9,** 601–614 (1990).
32. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30,** 2785–2791 (2009).
33. Schüttelkopf, A. W. & van Aalten, D. M. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr. D* **60,** 1355–1363 (2004).

LETTER

doi:10.1038/nature11013

Systematic discovery of structural elements governing stability of mammalian messenger RNAs

Hani Goodarzi[1,2]†, Hamed S. Najafabadi[3,4]†, Panos Oikonomou[1,2]†, Todd M. Greco[2], Lisa Fish[5], Reza Salavati[3,4,6], Ileana M. Cristea[2] & Saeed Tavazoie[1,2]†

Decoding post-transcriptional regulatory programs in RNA is a critical step towards the larger goal of developing predictive dynamical models of cellular behaviour. Despite recent efforts[1–3], the vast landscape of RNA regulatory elements remains largely uncharacterized. A long-standing obstacle is the contribution of local RNA secondary structure to the definition of interaction partners in a variety of regulatory contexts, including—but not limited to—transcript stability[3], alternative splicing[4] and localization[3]. There are many documented instances where the presence of a structural regulatory element dictates alternative splicing patterns (for example, human cardiac troponin T) or affects other aspects of RNA biology[5]. Thus, a full characterization of post-transcriptional regulatory programs requires capturing information provided by both local secondary structures and the underlying sequence[3,6]. Here we present a computational framework based on context-free grammars[3,7] and mutual information[2] that systematically explores the immense space of small structural elements and reveals motifs that are significantly informative of genome-wide measurements of RNA behaviour. By applying this framework to genome-wide human mRNA stability data, we reveal eight highly significant elements with substantial structural information, for the strongest of which we show a major role in global mRNA regulation. Through biochemistry, mass spectrometry and *in vivo* binding studies, we identified human HNRPA2B1 (heterogeneous nuclear ribonucleoprotein A2/B1, also known as HNRNPA2B1) as the key regulator that binds this element and stabilizes a large number of its target genes. We created a global post-transcriptional regulatory map based on the identity of the discovered linear and structural *cis*-regulatory elements, their regulatory interactions and their target pathways. This approach could also be used to reveal the structural elements that modulate other aspects of RNA behaviour.

To isolate stability from other aspects of mRNA behaviour, we performed whole-genome mRNA stability measurements by incubating human MDA-MB-231 breast cancer cells in the presence of 4-thiouridine, which is efficiently incorporated into cellular RNA. Subsequently, 4-thiouridine-labelled transcripts were captured and quantified at different time-points after the removal of 4-thiouridine from the growth medium. We calculated a relative decay rate for each transcript based on the rate at which 4-thiouridine-labelled transcripts, in the absence of 4-thiouridine in the media, are replaced by newly synthesized unlabelled mRNAs in the population (Supplementary Fig. 1). These measurements were then used to identify the putative *cis*-regulatory elements (linear and structural) that underlie transcript stability. A number of methods have been previously introduced for discovering structural motifs mainly based on free energy minimization, local sequence alignments or a combination of both alignments and secondary structure predictions[3,6,8]. However, the extent to which these *in silico* predictions reflect stable *in vivo* molecular conformations has not been fully explored[9]. In fact, the RNA binding proteins and complexes that interact with their target transcripts may facilitate the formation of secondary structures *in vivo*. Thus, we sought to bypass the need for predicting thermodynamically stable secondary structures by efficiently enumerating a large space of potential structural motifs. We developed TEISER (Tool for Eliciting Informative Structural Elements in RNA), a framework for identifying the structural motifs that are informative of whole-genome measurements across all the transcripts. In this approach, structural motifs are defined in terms of context-free grammars[7] (CFGs) that represent hairpin structures as well as primary sequence information (see Methods and Supplementary Fig. 2). TEISER employs mutual information to measure the regulatory consequences of the presence or absence of each of roughly 100 million different seed CFGs (see Methods). Mutual information is a robust non-parametric measure that reveals general dependencies across discrete or continuous measurements[2,10]. For example, when applied to the transcript stability data, TEISER captures the dependency between the stability of each mRNA and the presence or absence of a given structural motif in its 5′ and 3′ untranslated regions (UTRs). TEISER, subsequently, uses these measurements to choose and further refine the most informative motifs, and performs a series of statistical tests—for example, randomization-based statistics and jackknifing tests—to achieve very low (<0.01) false-discovery rates (see Methods and Supplementary Fig. 2).

Application of TEISER to the mRNA stability measurements in MDA-MB-231 cells revealed eight strong structural motif predictions that passed our statistical tests aimed at finding the most likely elements causally involved in mRNA stability (Fig. 1 and Supplementary Fig. 3). Apart from being highly informative of mRNA stability measurements, these putative regulatory elements show a variety of other characteristics that support their functionality. For example, four of the discovered motifs are also informative of transcript stability measurements in mouse[11] (Supplementary Fig. 4a). Furthermore, these motifs are highly conserved between human and mouse genomes (see Methods and Supplementary Fig. 3) and are also informative of co-expression clusters discovered across independent whole-genome data sets (Supplementary Fig. 4b).

Among the putative structural motifs discovered by TEISER, we chose sRSM1 (structural RNA stability motif 1)—the most statistically significant 3′ UTR element (z-score = 122)—for further analysis. In order to probe the functionality of sRSM1 instances across the genome, we performed *in vivo* titration experiments using synthetic oligonucleotides[10,12]. Upon transfecting MDA-MB-231 cells with decoy RNA molecules harbouring sRSM1 instances (Supplementary Fig. 5), we observed a notable reduction in the level of endogenous transcripts that carried this motif, in comparison to their level in the control cells transfected with scrambled RNA molecules (Fig. 2). This global

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08540, USA. [2]Department of Molecular Biology, Princeton University, Princeton, New Jersey 08540, USA. [3]Institute of Parasitology, McGill University, Montreal, Quebec H3G1Y6, Canada. [4]McGill Centre for Bioinformatics, McGill University, Montreal, Quebec H3G1Y6, Canada. [5]Laboratory of Systems Cancer Biology, Rockefeller University, New York, New York 10065, USA. [6]Department of Biochemistry, McGill University, Montreal, Quebec H3G1Y6, Canada. †Present addresses: Department of Biochemistry and Molecular Biophysics, and Initiative in Systems Biology, Columbia University, New York, New York 10032, USA (H.G., P.O., S.T.); The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada (H.S.N.).

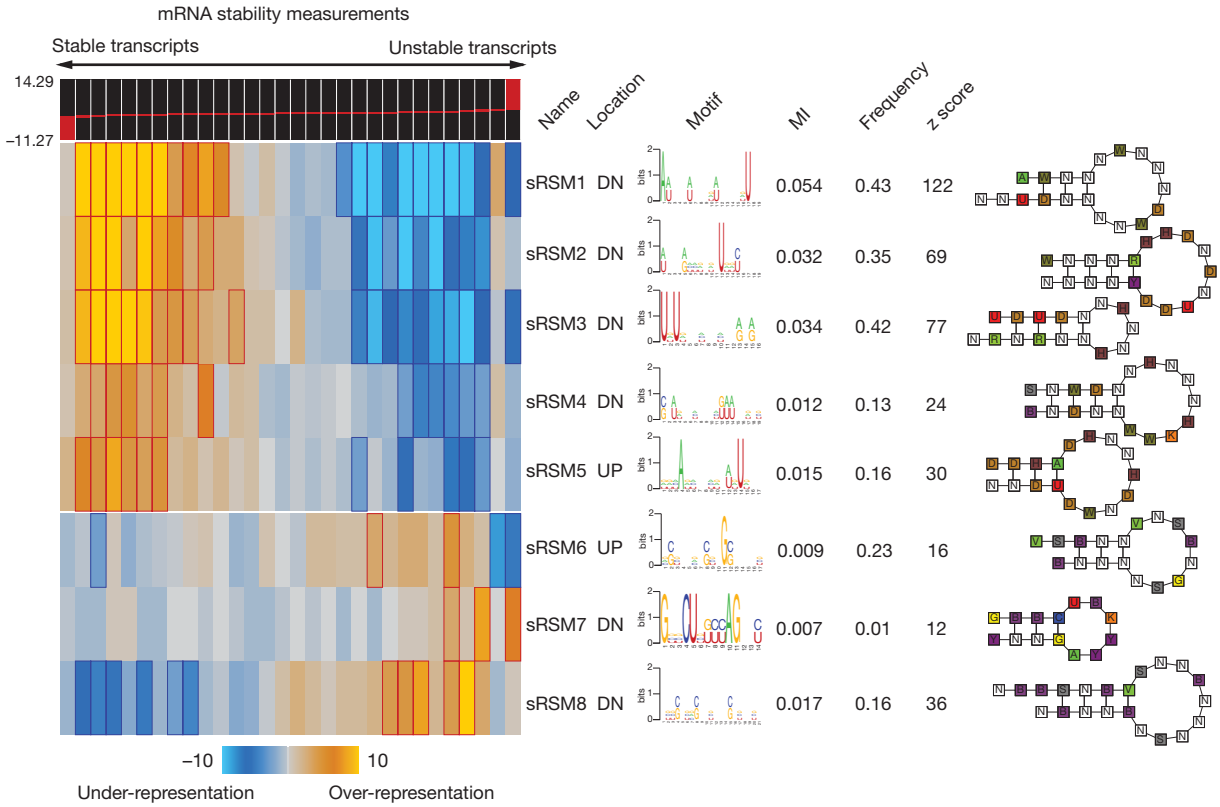


Figure 1 | Discovery of RNA structural motifs informative of genome-wide transcript stability. Each RNA structural motif is shown (far right) along with its pattern of enrichment/depletion across the range of mRNA stability measurements throughout the genome (far left). The panel labelled mRNA stability measurements shows how the transcripts are partitioned into equally populated bins based on their stability measures, going from left (highly stable) to right (unstable). In the heatmap representation, a gold entry marks the enrichment of the given motif in its corresponding stability bin (measured by log-transformed hypergeometric *P*-values), while a light-blue entry indicates motif depletion in the bin. Red and blue borders mark highly significant motif enrichments and depletions, respectively. From left to right, we show the motif names, their location (UP for 5′ UTR and DN for 3′ UTR), their sequence information ('motif', in the form of an alphanumeric plot), their associated mutual information values (MI; see below), their frequency (the fraction of transcripts that carry at least one instance of the motif), and their *z* score (see below). Each MI value is used to calculate a *z* score, which is the number of standard deviations of the actual MI relative to MIs calculated for 1.5 million randomly shuffled stability profiles. A structural illustration of each motif is also presented (far right) using the following single letter nucleotide code: Y = [UC], R = [AG], K = [UG], M = [AC], S = [GC], W = [AU], B = [GUC], D = [GAU], H = [ACU], V = [GCA] and N = any nucleotide.

downregulation points to the presence of a *trans*-acting factor that, upon interaction with sRSM1, stabilizes its target transcripts. The decoy (synthetic) sRSM1 elements compete with endogenous mRNAs for the putative *trans*-acting factor, which results in the observed reduction in the level of its target mRNAs. Furthermore, reporter constructs carrying instances of sRSM1 showed a marked decrease in transcript decay rate in comparison to scrambled controls, further suggesting a direct role for this structural element in transcript stability (Supplementary Fig. 6).

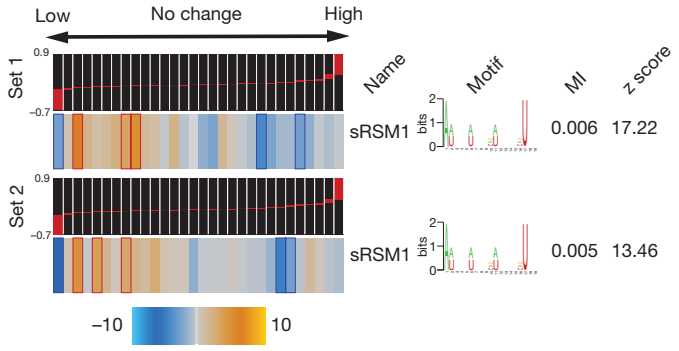


Figure 2 | The regulatory role of sRSM1. Whole-genome expression levels were measured in decoy-transfected samples relative to the controls transfected with scrambled RNA molecules (see Methods). The measurements were performed in duplicate, for two independent decoy/scrambled sets (the relative transcript levels were subsequently averaged across the two replicates in each set). Genes were sorted and quantized into equally populated bins based on the average log-ratio of their expression levels in the decoy samples relative to the scrambled controls. TEISER was used to show the enrichment/depletion patterns of transcripts harbouring sRSM1 in their 3′ UTRs. From left to right, we also show motif name, sequence, MI values and the associated *z* scores.

We used streptomycin-binding RNA aptamer immobilization coupled with mass spectrometry[13] to discover candidates that bind, *in vitro*, to the decoy instances of sRSM1, but not to the scrambled versions (Supplementary Fig. 7). After isolation under stringent conditions and in-solution digestion of RNA-bound proteins followed by nanoliquid chromatography-tandem mass spectrometry, we identified HNRPA2B1 as a promising candidate (Supplementary Table 1). This RNA-binding protein is a member of the A/B subfamily of heterogeneous nuclear ribonucleoproteins (hnRNPs)[14] and carries two repeats of quasi-RNA-recognition motif (qRRM) RNA binding domains (Supplementary Fig. 8). Moreover, the established roles of other members of this family, namely HNRNPD and HNRNA1, in regulating RNA stability[15] and binding terminal stem-loops[16] further suggest HNRPA2B1 as a functional regulator. Also, more than 4,000 transcripts carry potentially functional instances of sRSM1 (see Methods), implicating this motif as a major global regulator of mRNA stability. The HNRPA2B1 transcript, at the same time, is highly abundant in the cell (one standard deviation higher than average[17]), thus making it a promising candidate for global modulation of mRNA stability through sRSM1.

In order to directly assess the regulatory consequences of modulating HNRPA2B1, we performed knock-down experiments followed by gene expression profiling. Consistent with our prior observations, HNRPA2B1 knock-down caused a significant decrease in the expression level of transcripts carrying sRSM1 (Fig. 3a). Stability measurements in the knock-down cells confirmed that the observed downregulation of these transcripts was in fact due to changes in stability (see Methods), with the transcripts carrying sRSM1 elements showing a marked increase in their corresponding relative decay rates (Fig. 3b).

In principle, our observations are consistent with a possible indirect role for HNRPA2B1—brought about, for instance, by a common partner that binds both HNRPA2B1 and sRSM1 sites. The direct interaction between HNRPA2B1 and its potential target genes can be tested through cross-linking and immunoprecipitation of HNRPA2B1, which, through local ultraviolet photoreactivity of bases and amino acids, can detect direct physical interactions[18]. We expressed a tagged clone of HNRPA2B1 in MDA-MB-231 cells, and after ultraviolet-crosslinking, immunoprecipitated this protein and the target mRNA molecules that were bound to it. We then labelled the isolated RNA population and hybridized it to microarrays with the input total RNA as control (a method called RIP-chip[19]). We observed a highly significant enrichment of sRSM1 in the immunoprecipitated population (Fig. 3c). In order to reduce the background and better pinpoint the HNRPA2B1 binding sites, we treated the samples with nuclease before immunoprecipitation under denaturing conditions and sequenced the HNRPA2B1-bound RNA population (HITS-CLIP[20]). We observed that sRSM1 elements were significantly enriched in the identified putative binding sites, in comparison with randomly selected sequences[21] (Fig. 3d). These observations demonstrate that HNRPA2B1 directly

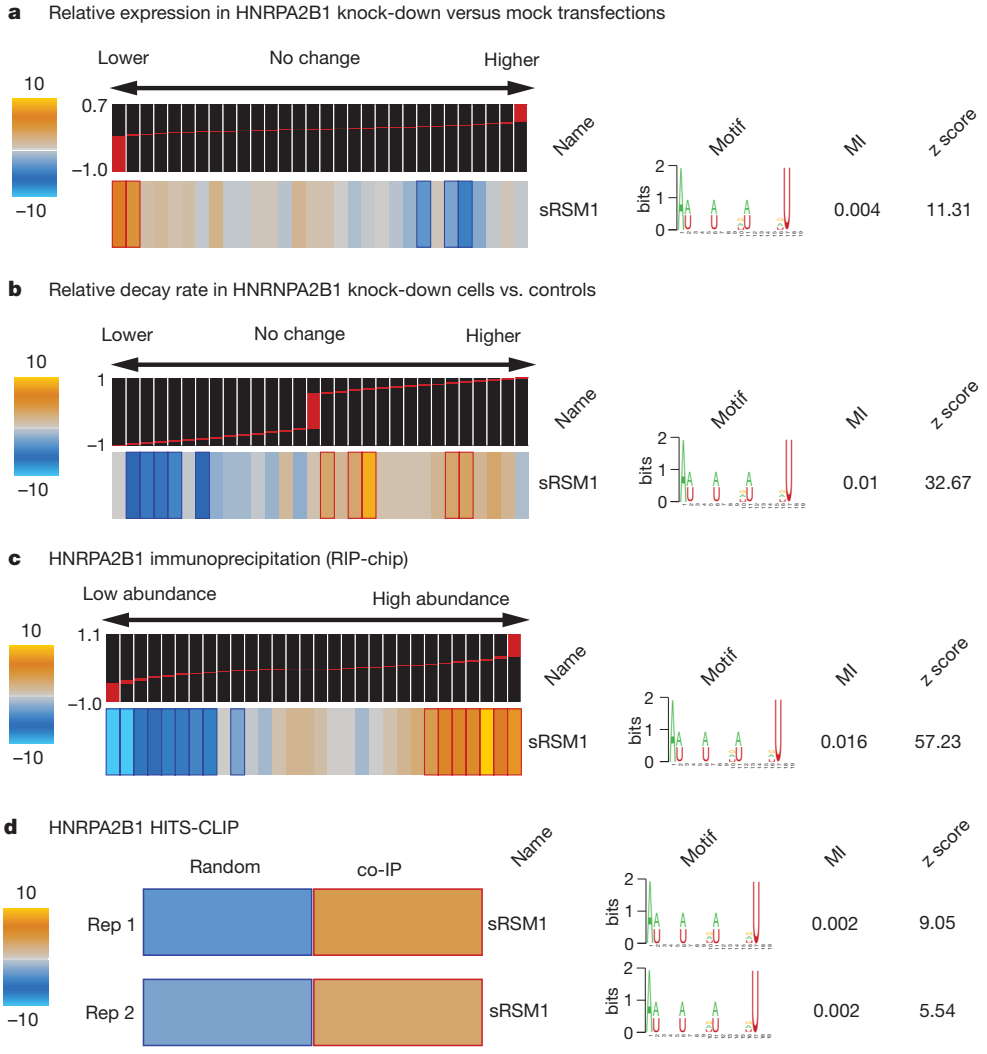


Figure 3 | HNRPA2B1 stabilizes transcripts through direct *in vivo* binding to sRSM1 structural motifs. a, Genome-wide expression levels were measured in HNRPA2B1 siRNA-transfected samples relative to mock-transfected controls. TEISER was used to capture the enrichment/depletion pattern of transcripts carrying sRSM1 across the relative expression values. Experiments were performed in triplicate, each with an independent siRNA targeting HNRPA2B1 and the resulting log ratios were averaged for each transcript. **b**, Transcript decay rates were compared in HNRPA2B1 knock-downs versus mock-transfected controls. These measurements were then analysed by TEISER to visualize the extent to which the decay rates of transcripts carrying sRSM1 elements were increased following HNRPA2B1 knock-down. **c**, Using ultraviolet-crosslinking followed by immunoprecipitation, mRNAs that bind HNRPA2B1 were extracted and compared against the input mRNA population (RIP-chip). The log ratio calculated for each mRNA denotes its abundance in the immunoprecipitated sample relative to the input control. Bins to the right contain the mRNAs that were captured as interacting partners with HNRPA2B1. Similar to the prior examples, TEISER was used to show the enrichment/depletion pattern of transcripts carrying sRSM1 in their 3′ UTRs. The values associated with each transcript were calculated as the average of log ratios from biological replicates. **d**, HNRPA2B1 binding sites were identified using immunoprecipitation followed by high-throughput sequencing (HITS-CLIP). Instances of the sRSM1 element are significantly enriched in these sites relative to a population of random sequences from 3′ UTRs that are not represented in the sequenced population.

interacts with sRSM1 *in vivo* and acts to stabilize its target transcripts through this regulatory element. These transcripts, in turn, modulate a variety of cellular processes and pathways. For example, we observed a significant positive correlation between sRSM1 target transcripts and doubling-time in NCI-60 breast cancer cell lines (Fig. 4a). Indeed, knocking-down HNRPA2B1 resulted in a slight but significant increase in growth rate (by 10%, *P*-value $<10^{-8}$), further highlighting the regulatory role of this global modulator in a key cellular process (Fig. 4b).

Revealing the detailed post-transcriptional regulatory code relies on the discovery of all the *cis*-regulatory elements that contribute to changes in transcript abundance. In addition to the sRSMs identified through TEISER, we also discovered a large diverse set of lRSMs (linear RNA stability motifs), including six known microRNA recognition sites, that are informative of transcript stability measurements (Supplementary Fig. 9). These motifs were identified by FIRE[2] (Finding Informative Regulatory Elements), a framework for discovering informative linear motifs. Combining these two approaches provided us with an extensive set of putative regulatory elements that cover both structural and primary sequence components. The next step in deciphering the post-transcriptional regulatory program involves the identification of target pathways that are potentially modulated by each element. Using iPAGE[10] (Pathway Analysis of Gene Expression), we showed that our discovered elements probably target a diverse array of cellular processes and pathways (Supplementary Fig. 10). For example, the sRSM1 structural element is significantly enriched in the 3′ UTRs of the genes involved in 'Notch signalling', while avoiding the UTRs of other pathways such as 'nucleosome assembly' (Supplementary Fig. 11). These results demonstrate that while post-transcriptional regulatory mechanisms are poorly characterized, they have potentially far-reaching impact on specific cellular processes.

Regulatory programs often employ combinatorial interactions between various *cis*-regulatory elements to modulate gene expression[2,22]. We used mutual information to reveal such potential interactions in the post-transcriptional regulatory programs governing mRNA stability (Supplementary Figs 12 and 13). For example, sRSM1 showed significant interactions with a number of structural and linear motifs, including sRSM8 and sRSM3 (Supplementary Fig. 11). These observed interactions might reflect cross-talk, or insulation, between the underlying regulatory processes that act upstream of these elements. The full map of such interactions (Supplementary Figs 14 and 15) reveals a complex network of motif-pathway relationships that set the stage for molecular dissection and predictive modelling of post-transcriptional regulation from sequence.

Whereas we have studied mRNA stability under normal and static conditions in a single cell line, the full regulatory program that governs mRNA stability is likely to involve a much richer repertoire of *cis*-regulatory elements operating within a more complex regulatory network. Also, although we have focused on transcript stability, our framework is general in concept and can be employed to study regulatory programs governing other aspects of RNA biology. For example, the established role of local secondary structures in shaping the splicing code[4,23] suggests alternative splicing as a prominent area for analysis using this framework. The large repertoire of publicly available whole-genome expression data sets similarly offers a rich resource for identifying the post-transcriptional regulatory modules that underlie steady-state measurements.

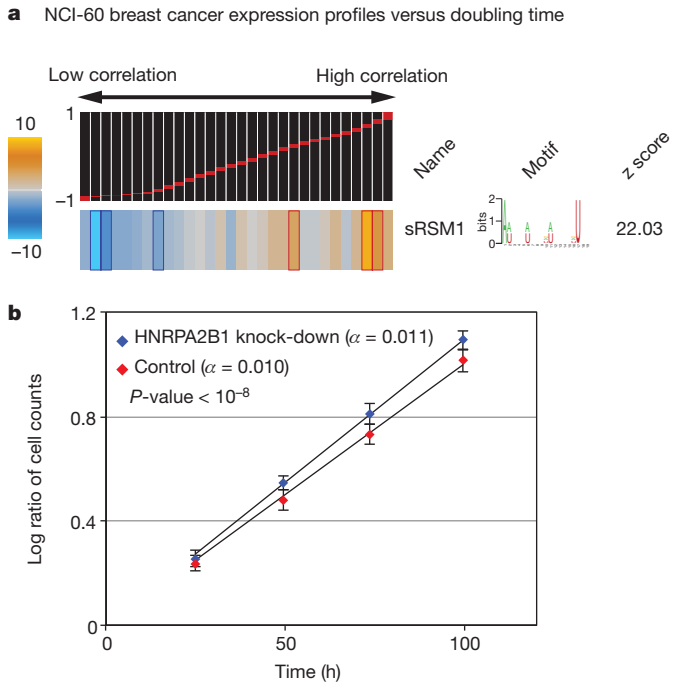


Figure 4 | HNRPA2B1 regulates growth rate. a, Whole genome expression levels across five breast cancer cell lines (MCF7, MDA-MB-231, HS578T, BT-549 and T47D) were correlated against their doubling times[17]. The resulting values, ranging from −1 to 1, were analysed by TEISER to probe the enrichment/depletion pattern of transcripts carrying sRSM1. **b**, The growth of HNRPA2B1 siRNA-transfected samples was compared to those of mock-transfected controls. For each time-point, the number of cells in four independent samples was counted in duplicates ($n = 8$), yielding an estimated growth-rate (α). Shown are the average log-ratios, their standard deviation at each time-point, and the statistical significance of the observed difference in growth-rate.

METHODS SUMMARY

TEISER relies on calculating mutual information (MI) values between whole-genome measurements and millions of predefined structural motifs. The statistically significant motifs are then optimized and elongated through a heuristic search algorithm. The mRNA stability measurements were performed using a previously published method[1]. The decoy/scrambled experiments and siRNA knock-downs were performed using lipofectamin 2000 reagent (Invitrogen). For hybridizations, we used human 4 × 44k whole-genome human arrays (Agilent). Isolation and identification of RNA-binding proteins were based on previously published protocols[13,24]. HNRPA2B1 target transcripts were isolated based on the CLIP protocol[18].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 10 August 2011; accepted 2 March 2012.
Published online 8 April 2012.

1. Dölken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14,** 1959–1972 (2008).
2. Elemento, O., Slonim, N. & Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* **28,** 337–350 (2007).
3. Rabani, M., Kertesz, M. & Segal, E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl Acad. Sci. USA* **105,** 14885–14890 (2008).
4. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).
5. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nature Rev. Genet.* **12,** 641–655 (2011).
6. Pavesi, G., Mauri, G., Stefani, M. & Pesole, G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucleic Acids Res.* **32,** 3258–3269 (2004).
7. Searls, D. B. The language of genes. *Nature* **420,** 211–217 (2002).
8. Hofacker, I. L., Fekete, M. & Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319,** 1059–1066 (2002).
9. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467,** 103–107 (2010).
10. Goodarzi, H., Elemento, O. & Tavazoie, S. Revealing global regulatory perturbations across human cancers. *Mol. Cell* **36,** 900–911 (2009).
11. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
12. Cutroneo, K. R. & Ehrlich, H. Silencing or knocking out eukaryotic gene expression by oligodeoxynucleotide decoys. *Crit. Rev. Eukaryot. Gene Expr.* **16,** 23–30 (2006).

13. Windbichler, N. & Schroeder, R. Isolation of specific RNA-binding proteins using the streptomycin-binding RNA aptamer. *Nature Protocols* **1,** 637–640 (2006).
14. Biamonti, G., Ruggiu, M., Saccone, S., Della Valle, G. & Riva, S. Two homologous genes, originated by duplication, encode the human hnRNP proteins A2 and A1. *Nucleic Acids Res.* **22,** 1996–2002 (1994).
15. Wilusz, C. J., Wormington, M. & Peltz, S. W. The cap-to-tail guide to mRNA turnover. *Nature Rev. Mol. Cell Biol.* **2,** 237–246 (2001).
16. Michlewski, G. & Caceres, J. F. Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. *Nature Struct. Mol. Biol.* **17,** 1011–1018 (2010).
17. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24,** 227–235 (2000).
18. Jensen, K. B. & Darnell, R. B. CLIP: crosslinking and immunoprecipitation of *in vivo* RNA targets of RNA-binding proteins. *Methods Mol. Biol.* **488,** 85–98 (2008).
19. Keene, J. D., Komisarow, J. M. & Friedersdorf, M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Protocols* **1,** 302–307 (2006).
20. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456,** 464–469 (2008).
21. Giannopoulou, E. G. & Elemento, O. An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics* **12,** 277–294 (2011).
22. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117,** 185–198 (2004).
23. Yang, Y. *et al.* RNA secondary structure in mutually exclusive splicing. *Nature Struct. Mol. Biol.* **18,** 159–168 (2011).
24. Greco, T. M., Yu, F., Guise, A. J. & Cristea, I. M. Nuclear import of histone deacetylase 5 by requisite nuclear localization signal phosphorylation. *Mol. Cell Proteomics* **10,** M110.004317 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the members of the Tavazoie laboratory for comments on the project and manuscript. We are also grateful to N. Pencheva, B. Tsui, S. Tavazoie and L. Dölken for their intellectual and technical contributions. L.F. was supported by a Ruth L. Kirschstein National Research Service Award (T32-GM066699). S.T. was supported by grants from NHGRI (2R01HG003219) and the NIH Director's Pioneer Award.

Author Contributions H.G., H.S.N. and S.T. conceived and designed the study. H.G. and H.S.N. developed TEISER. R.S. contributed to the execution of the study. H.G., H.S.N., T.M.G., P.O., I.M.C. and S.T. designed the experiments. H.G., P.O., L.F. and T.M.G. performed the experiments. H.G., H.S.N. and T.M.G. analysed the results. H.G., H.S.N. and S.T. wrote the paper.

Author Information The microarray and high-throughput sequencing data are deposited at GEO under the umbrella accession number GSE35800. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.T. (st2744@columbia.edu).

METHODS

TEISER: detailed description of the algorithm. Genome profile. A genome profile is defined across the genes in the genome, where each gene is associated with a unique measurement. Whole-genome measurements, discrete or continuous, can be obtained from a variety of experimental or computational sources (for example, Supplementary Fig. 1).

Structural motif definition. Each structural motif is defined as a series of context-free statements that define the structure and sequence of the motif (Supplementary Fig. 2). A context-free grammar is a set of production rules that describes how phrases are made from their building blocks. Considering a structured RNA molecule as a phrase, its potential building blocks are the different base pairs and bulges. Loops can be considered as bulges that happen at the beginning of phrases. Also, internal loops can be considered as combination of left and right bulges in the middle of phrases. The context-free grammar that we have used contains the following production rules: $S \rightarrow S$[AUCGN], $S \rightarrow$[AUCGN]S, $S \rightarrow$[AUCGN]S[AUCGN]; wherein the first production rule depicts a right bulge, the second production rule results in a left bulge, and the third production rule creates a base-pairing. For example, consider the stem loop AAACGCUUU (the stem region is underlined). Let the symbol S be a non-terminal symbol that stands for this stem loop; the production rule $S \rightarrow SG$ adds a G to the 3′ end of the molecule, creating a new S, AAACGCUUUG, which has an unpaired 3′-end G. Next, using the production rule $S \rightarrow GSC$, we can add a G to the 5′ end and a C to the 3′ end of the molecule and make them pair with each other, again creating a new S, GAAACGCUUUGC, which can be further expanded in this way. Note that the G that we added in the previous step has now become a right bulge.

Motif profile. For every given motif, we create a binary vector across all the genes, in which '1' denotes the presence and '0' denotes the absence of that motif. This vector is called a motif profile.

Creating seed CFGs. We used, as the seed motifs, an exhaustive set of context-free statements that represented all possible stem-loop structures that satisfied the following criteria: stem length of at least 4 bp and at most 7 bp; loop length of at least 4 nt and at most 9 nt; at least 4 and at most 6 production rules representing non-degenerate bases (that is, production rules that are not $S \rightarrow SN$, $S \rightarrow NS$, or $S \rightarrow NSN$); and information content of at least 14 bits and at most 20 bits. The information content of the motif M, which is represented by n production rules, was defined as $-\log_2(p_M)$, wherein p_M is the probability that a random sequence of length l matches the n production rules of motif M, with l being equal to $2 \times n_1 + n_2$ in which n_1 is the number of production rules that represent base pairings and n_2 is the number of production rules that represent bulges ($n_1 + n_2 = n$).

Quantizing continuous genome profiles. Mutual information is defined for both continuous and discrete random variables; however, in practice, continuous data are discretized before calculating the mutual information (MI) values. Our quantization procedure involves using equally populated 'bins'. Thus, the discretization step only requires a single parameter, that is, the number of genes in each bin. In TEISER, we have set the default number of bins to 30 ($N_e = 30$). It should be noted that the results are not sensitive to variations in the value of N_e as long as N_e is >10 and each bin has more than $\sim$100 associated transcripts.

Removing recently duplicated genes. Recently duplicated members of gene families or transposons often share a significant amount of sequence identity in their UTRs. They also tend to cross-hybridize on the arrays and show a high artificial correlation. This would in turn bias our search towards conserved elements in the UTRs of these genes. In TEISER, similar to FIRE[2], we remove the duplicates that have similar values (for example, fall in the same bin after quantization of the input genome profile). A MegaBlast E-value cutoff of 1×10^{-15} was used to identify duplicates.

Calculating the mutual information values. We performed mutual information (MI) calculations between the genome profile and the motif profiles using algorithms introduced and described elsewhere[2,10]. These algorithms take the necessary steps to ensure reliable MI calculations (for example, minimum sample sizes for reliable estimation of joint distributions).

Randomization-based statistical testing. To assess the statistical significance of the calculated MI values, TEISER uses a non-parametric randomization-based statistical test. In this test, the genome profile is shuffled 1,500,000 times and the corresponding MI values are calculated. A motif is deemed significant only if the real MI value is greater than all of the randomly generated ones. In TEISER, in order to minimize the required number of tests, structural motifs are first sorted based on the MI values (from high to low) and the statistical test is applied in order. When 20 contiguous motifs in the sorted list do not pass the test, the procedure is terminated.

Optimization of the identified seeds into more informative motifs. Our initial collection of structural motifs, despite being large, is a coarse-grained sampling of the entire space. Mainly, it provides us with a set of informative seeds that should be later optimized into closer representations of their actual form[2]. Accordingly, all the structural motifs that pass the previous stage are further optimized and elongated. The process involves: (1) optimization: randomly choose one of the context-free statements (production rules) from the motif and convert its sequence information to all possible combinations of nucleotides. Evaluate all the resulting structural motifs and accept the one that results in the highest MI value. (2) Elongation: production rules are added to the end of the context-free phrase that represents the motif, thus extending its effective length in the form of a base pair or a bulge. The increase in length is similarly accepted only if it results in a higher MI value.

Removing redundantly informative structural motifs. Motifs that redundantly represent the same potential *cis*-regulatory elements are identified and removed using the concept of conditional information as described before[2,10].

Finding robust motifs. TEISER also performs jack-knife resampling to find robust motifs that are not over-sensitive to the composition of the input data. For each predicted motif, we perform 10 jackknifing trials where, in each trial, one third of the genes are randomly removed and the mutual information value and its statistical significance is evaluated. The robustness score is then defined as the number of trials in which the motif remains significant (scores better in the original genome profile than in all the randomly shuffled genome profiles) after resampling, ranging from 0/10 to 10/10. By default, TEISER requires the motif to be significant in more than half of the trials (a robustness score equal to or greater than 6/10). While this parameter can be changed at the user's discretion, our experience with both TEISER and FIRE[2] suggests that this threshold results in very low false discovery rates across a variety of data sets (discrete and continuous).

Patterns of motif enrichment and depletion. For a given motif, a high mutual information value results from the non-random distribution of its targets across the input range. This results in significant patterns of enrichment and depletions across the genome profile, which can be quantified by calculating enrichment/depletion scores. These scores result from the log transformation of P-values calculated based on the hypergeometric distribution, as described previously[2].

Final statistical tests. In case the genome profile is continuous, one can require TEISER to return motifs that are enriched at one end of the data range or the other (for example, structural motifs in Fig. 1). TEISER accomplishes this through calculating the Spearman correlation between the enrichment scores and the average data value across all the bins. For the structural motifs in Fig. 1, the P-value threshold for these Spearman correlations was set to 0.001 (for Supplementary Fig. 3, this value is 0.01 which puts the FDR at 10%). It should be noted, however, that other statistical tests could be used in this step at the discretion of the user. The goal, ultimately, is to identify the motifs that show significant enrichments at either end of the data range.

Inter-species conservation. For each motif, we also calculate a conservation score based on its network-level conservation with respect to a related genome[2]. For this, orthologous transcripts in both genomes are scanned for the presence/absence of the motif. The overlap of positive sequences between the orthologous sequences is used to calculate a hypergeometric P-value[2]. The conservation score is then defined as $1 - P$, which ranges between 0 and 1 (1 being highly conserved between the two genomes). In this study, we have used the human and mouse genomes to calculate the conservation scores associated with each structural motif.

Finding potentially active instances of each motif. As described previously[2], we defined the target genes of a predicted motif as all transcripts whose 3′ or 5′ UTRs contain the motif and are associated with a category/bin where the motif is enriched. In other words, these are the transcripts whose UTRs contain potentially 'active' motif occurrences. Upon identifying these likely targets for each structural motif, a weight-matrix can be generated from these potentially functional instances as a post-processing step (Supplementary Table 2).

False-discovery rate. In order to assess the false discovery rate, we ran 30 trials with shuffled 5′ and 3′ UTR sequences. In all the trials, not a single motif passed all the statistical tests. Thus, in case of the stability data set, the number of false positives in each trial, on average, is smaller than $1/30 \approx 0.34$, which corresponds to an FDR of <0.01.

Predicting functional interactions. Given two motifs, structural or linear, one can assess their putative functional interaction through measuring how informative the presence of one would be about the presence or absence of the other. For revealing these interactions, we again use mutual information values calculated for pairwise motif profiles of structural and linear motifs. Randomization-based statistical tests are then used to find the significant interactions. For this, one of the motif profiles is shuffled 10,000 times and the interaction is deemed significant only if the real mutual information value is higher than all the 10,000 random ones.

Predicting the target pathways. iPAGE[10], with default settings, was used to identify the likely pathways that are regulated by the discovered structural and linear motifs.

Availability. TEISER is available online for download at https://tavazoielab.c2b2.columbia.edu/TEISER.

Measuring mRNA stability. RNA stability measurements were performed based on a previously published protocol[1]. In short, MDA-MB-231 cells at 70% confluency were incubated in the presence of 25 μM 4-thiouridine (Sigma) for 4 h. Then the cells were washed with fresh media (DMEM + 10% FBS) and incubated for 0, 1, 2 and 4 h. At each time point, cells were washed with cold PBS and RNA extraction was performed using a total RNA purification kit (Norgen Biotek). The 4-thiouridine thiol groups were then biotinylated using EZ-Link Biotin-HPDP (Pierce). We subsequently used μMacs magnetic columns (Miltenyi Biotec) to capture the labelled RNAs. The resulting samples were then processed for one-colour hybridization using a one-colour low-input quick-amp labelling kit (Agilent) and hybridized according to the manufacture's instructions. A one-colour RNA spike-in kit (Agilent) was used as endogenous control to normalize values between arrays. For each transcript, the drop in signal as a function of time was used as a measure of mRNA stability (Supplementary Fig. 1): $r = -ln\left(\frac{S_t}{S_0}\right) \Big/ t$, where S_t denotes signal at time t. Linear regression was used to calculate r for each transcript based on the hybridization signals from the four time points. It should be noted that TEISER is a non-parametric approach, thus it is the ranking rather than the actual stability values that underlies our motif discovery.

Transfection of decoy and scrambled oligonucleotides. We chose real instances of the sRSM1 structural motifs from NM_014363, which contains four instances of sRSM1, to create two decoy sets of sequences, each containing two of these instances (underlined) along with part of the real sequences as context. Set 1: AAAACTATTTTGAAGATGGTGGTGAGCTGCAAAATAGCTGGATGGATTTGAATGATTGGGATGATACATCATTGAACTGCACTTTATATAACCAAAGCTTAGCAGTTTGTTAGATAAGAGTCTATGTATGTCTCTGGTTAGGATGAAGTTAATTTTATGTTTTTAACATGGTATTTTTGAAGGAGCTAATGAAACACTGG. Set2: ATTGTTTCTGGAAACTGCTTGCCAAGACAACATTTATTAACTGTTAGAACACTTGCTTTATGTTTGTGTGTACATATTTTCCACAAATGTTATAATTTATATAGTGTGGTTGAACAGGATGCAATCTTTTGTTGTCTAAAGGTGCTGCAGTTAAAAAAAAAAACAACCTTTTCTTTCAATATGGCATGTAGTGGAGTTTTT. For the scrambled controls, we used the shuffled version of the putative binding sites (see Supplementary Fig. 5). These two decoy/scrambled sets were then chemically synthesized (IDT). An upstream T7 promoter was used to transcribe the constructs *in vitro* using Megascript T7 kit (Ambion). In order to reduce cytotoxicity, RNA molecules were capped and poly-A tailed using Cap Analogue (Ambion) and poly-A polymerase (NEB). MDA-MB-231 cells at 80% confluency were transfected with the resulting RNA oligos using Lipofectamin 2000 reagent (Invitrogen) according to manufacturer's recommendations. Experiments were performed in duplicates for each set. Forty-eight hours post-transfection, we extracted RNA and differentially labelled the samples with Cy3 or Cy5 dyes. The samples were then hybridized on Agilent human gene expression arrays (4 × 44k). The Cy3/Cy5 ratios from the two biological replicates were then averaged into a single data set as log of ratios, which was then analysed by TEISER.

Reporter system for testing the functionality of sRSM1 instances. The plasmid pcDNA5/FRT/TOPO (Invitogen) was used to clone a GFP-coding sequence along with a gateway cloning site downstream of GFP (in its 3′ UTR). Decoy and scrambled sequences (Set 1 in the previous section) were subsequently cloned into the resulting construct using the gateway site. The resulting plasmids were transfected into the Flp-In 293 cell line (Invitrogen), and the cells were grown in Hygromycin for selecting stably transfected cells. The resulting cell lines, named Flp-In 293 GFP-Decoy and Flp-In 293 GFP-Shuffled, were subjected to FACS measurements to quantify GFP expression. For the decay rate measurements, cells were incubated in media with 5 μg ml^{-1} of α-amanitine (Sigma). Time points were taken at 0, 1.5, 3 and 6 h in duplicates for Flp-In 293 GFP-Decoy and Flp-In 293 GFP-Shuffled cells. Quantitative PCR (Fast SYBR Green Master Mix, Ambion) was then used to determine the relative quantity of GFP transcript in each cell line at different time-points using 18S rRNA as endogenous control.

Identifying binding candidates of sRSM1. We used a published protocol[13] to isolate potential RNA-binding proteins that bind sRSM1. In short, the StreptoTag aptamer was added downstream of the Set 1 decoy and scrambled sequences. The resulting RNAs were then immobilized on a dihydrostreptomycin Sepharose column (GE Healthcare) and were used to immunoprecipitate potential partners. Total protein was extracted from MDA-MB-231 cells (Total Protein Extraction Kit, Millipore), 1,000 μg of which was used as input to each column. Samples were then washed, eluted in 10 μM streptomycin and subjected to in-solution digestion[24,25]. Tryptic peptides were then analysed by nanoliquid chromatography-tandem mass spectrometry using an Ultimate 3000 nRSLC (Dionex) coupled online to an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific), as previously described[24].

HNRPA2B1 knock-down. The ON-Target*plus* (Dharmacon) set of siRNAs for HNRPA2B1 (target sequences: GAGGAGGAUCUGAUGGAUA, GGAGAGUAGUUGAGCCAAA, and GCUGUUUGUUGGCGGAAUU) were used to transfect MDA-MB-231 cells (grown in D10F medium) using Lipofectamine 2000 (Invitrogen). Three of the four tested siRNAs resulted in a substantial knock-down in HNRPA2B1 (more than twofold reduction in expression, log ratio >0.4 and $P < 10^{-7}$) and their corresponding samples were used for hybridization. Forty-eight hours post-transfection, we extracted total RNA from each sample along with mock-transfected controls. We then differentially labelled the RNA samples with Cy3 and Cy5 dyes and hybridized them to Agilent human gene expression arrays (4 × 44k). The log of signal ratios was used as a measure of differential expression between the samples and controls. These values were averaged across the three samples and were subsequently analysed by TEISER to assess the enrichment/depletion pattern of sRSM1 across the distribution.

For the decay rate measurements, forty-eight hours post-transfection, cells were incubated in media with 5 μg ml^{-1} of α-amanitine (Sigma). Time points were taken at 0, 1, 2 and 4 h in duplicates for the siRNA-transfected samples and mock-transfected controls. Each sample was then Cy3-labelled and hybridized to expression arrays (Agilent 4 × 44k) in duplicates and the reported signals were used to calculate decay rates. Following this procedure, for each transcript, four decay rates (two biological replicates, each having two technical replicates) were calculated from the siRNA-transfected samples and four decay rates from the controls. For each transcript, we then calculated a value according to $s(1 - P)$, where P is the t-test P-value between the two sets and s denotes whether the decay rates are higher in the siRNA samples (+1) or the mock controls (−1). After this transformation, the data range is between −1 and 1 with the background genes (the transcripts that show little change between the two samples) around 0. TEISER was then used to visualize the enrichment pattern of sRSM1 across this data range.

Identifying transcripts that interact with HNRPA2B1 (RIP-chip). A myc-tagged ORF clone of HNRPA2B1 (variant A2, OriGene) was transfected into MDA-MB-231 cells (grown in D10F medium) using Lipofectamine LTX and Plus reagent (Invitrogen). Seventy-two hours post-transfection, the cells were washed with cold PBS and ultraviolet-irradiated at 4,000 mJ cm^{-2}. The cells were then collected and lysed with 1 ml M-PER Reagent (Pierce) and 10 μl RNasin (NEB). The samples were subjected to DNase treatment (baseline ZERO DNase) for 15 min at 37 °C. Samples were then centrifuged at 16,000*g* at 4 °C for 20 min to pellet the cell debris. Immunoprecipitation of tagged HNRPA2B1 protein was performed using Mammalian c-Myc Tag IP/Co-IP Kit (Pierce) per manufacturer's instructions. Upon elution, samples were subjected to proteinase K digestion and polyadenylation. The RNA molecules in each sample were extracted using RNeasy MinElute Cleanup Kit (Qiagen) and Cy3-labelled using low-input quick-amp labelling kit (Agilent). As control, we used Cy5-labelled RNA samples extracted before HNRPA2B1 immunoprecipitation. The samples were hybridized to Agilent human gene expression arrays (4 × 44k) and the log of signal ratios was used as a measure of transcript affinity to HNRPA2B1. For each transcript, affinity values were averaged across two biological replicates and TEISER was used to assess the enrichment/depletion pattern of sRSM1.

Identifying 3′UTR binding sites of HNRPA2B1 (HITS-CLIP). A strategy similar to that of target transcript identification was used to discover the HNRPA2B1 binding sites. Upon ultraviolet-irradiation of mycHNRPA2B1-transfected cells, the samples were subjected to the HITS-CLIP protocol previously described elsewhere[26]. ChIPSeeqer[21], an integrated ChIP-seq analysis platform, was used to identify binding sites and extract real and random sequences (default parameters) for analysis with TEISER.

Measuring growth-rates in HNRPA2B1 knock-down cells. HNRPA2B1 siRNAs (Dharmacon) were used to knock-down the expression of this regulator. Seventy-two hours post-transfection, four independent samples were harvested and counted in duplicates as the baseline number of cells at time zero. Similarly, samples were counted at 25, 49.5, 73.5 and 99.5 h time-points. The same experiment was performed for mock-transfected cells. Using an exponential growth model, the log-ratio of the counted cells at each time-point was used to estimate a growth rate for siRNA-transfected and mock-transfected samples. ANCOVA was used to determine the P-value associated with the observed differences between the two growth rates.

25. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nature Methods* **6,** 359–362 (2009).
26. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460,** 479–486 (2009).

CAREERS

EUROPE Student group calls for more funding to support mobility **p.270**

INNOVATION US model is best for commercialization of inventions **p.270**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

E. SARINENA

COLUMN

Enterprising science

Graduate students and postdocs are often best placed to turn basic research into entrepreneurial gold, argues **Peter Fiske**.

Until recently, entrepreneurship was relatively unheard of as a career path for graduate students and postdocs. Early-stage researchers were expected to remain in the laboratory, where producing results and publications was their priority. If the resultant technology had commercial potential, it was their supervisor who disclosed the invention to the university to patent and license, or left the university to start a company themselves. However, the traditional model for technology commercialization and entrepreneurship in universities is flawed in two ways. First, universities find that simply licensing patents to private companies yields poor results: either the licence fee is too high for the company, or the business is ill-equipped to fill the technological gap between the idea and its commercial success. Even the 1980 Bayh–Dole Act in the United States, which provided laws to facilitate commercialization, didn't help to generate the incentive and the tools needed.

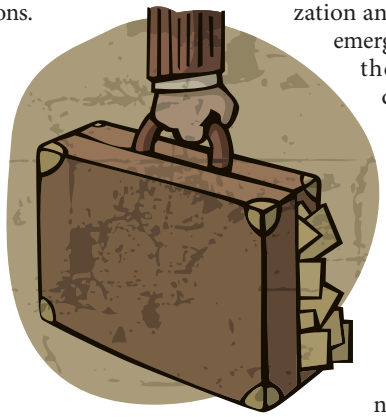
Second, supervisors are not the best people to start companies; they already have a job and do not have the time needed to start up a business. Although some principal investigators have an extraordinary track record of developing innovative technologies into thriving commercial businesses, the motivation for many is tempered by the reality that to embark on a risky entrepreneurial venture would come at the expense of an academic career.

However, that is not the case for ▶

EUROPE

Mobility boost

The European Student Union (ESU), which represents more than 11 million students, is calling for measures to boost student movement in Europe. A statement on 26 April from the union in Brussels says that a voluntary agreement, known as the Bologna Process, has stalled. Improving mobility was a motivation for the process, which was adopted in 1999 to make quality-assurance standards comparable across the participating nations. For better mobility, more graduate-level grants need to become portable, says ESU chairman Allan Päll. Portable schemes, such as Marie Curie Actions and Erasmus, represent only a tiny percentage of the available funding, he notes.



A. ZABELLA

CAREER DEVELOPMENT

Best practice

Recruitment is now more transparent and participation in performance reviews more common for UK academic researchers, finds a publicly funded group analysing the impact of academic standards. Vitae, a research-career advisory organization in Cambridge, UK, reviewed the success of a 2008 voluntary concordat among universities to improve the work life of researchers. The agreement aims to boost the appeal of research careers by setting guidelines for support. Further progress is needed, says Vitae chair Janet Metcalfe, including the greater engagement of researchers in their career development; Vitae are due to launch an online career-tool this autumn for researchers.

INNOVATION

More US success

An invention by an academic in the United States has a better chance of going to market than it does in other nations, a study finds. *University Entrepreneurship and Professor Privilege*, a working paper released by the Research Institute of Industrial Economics in Stockholm on 12 April, also finds that faculty members in other nations are more likely to try to launch their own inventions. Co-author Erika Färnstrand Damsgaard, a research fellow at the institute, says that US technology-transfer offices have more market-analysis skills, invest more in commercialization and often license to solid businesses, boosting the chances of success.

▶ early-stage researchers. Although they too may be unfamiliar with the process of starting up a company, they have a powerful incentive: they need a job. Young scientists should recognize the big contribution that they can make in spearheading the entrepreneurial process, and principal investigators should recognize the part that their postdocs and graduate students can play.

TAKING THE LEAD

Early-stage researchers are at the centre of a new model of technology commercialization and entrepreneurship that is emerging across the United States; they are the ones who are developing business plans and starting companies. This role reversal makes sense. Graduate students and postdocs are in the lab actually doing the experiments that could lead to a new technology, so they often have the best insight into the details of the invention and what is needed to improve it. Unlike principal investigators, who have to split their time and attention across numerous research projects and staff, early-stage researchers can focus all of their time and attention on one project. This single-minded focus is crucial for nurturing an invention. Rather than being just a pair of hands in the lab, graduate students and postdocs are now directing the commercialization effort, with the principal investigator taking on an advisory role.

As well as being familiar with the technology, graduate students and postdocs are often highly motivated, extremely resourceful and adaptable — all perfect attributes for being successful in a small, emerging entrepreneurial venture. No wonder they make the perfect 'number one' at a start-up.

This is not to say that supervisors no longer have an important role. The principal investigator's name is often attached to an innovative piece of technology that has been developed in their lab. In many cases, this is entirely appropriate because the principal investigator may have had the initial idea. And having a well-known name associated with the technology can be an advantage for a start-up company trying to secure investment.

Early-stage scientists have more start-up help at their disposal than ever before. Some of the leading research universities are recognizing the realities of this new model of technology commercialization, and are starting to put into place an innovation ecosystem — a term often used by universities and the US National Science Foundation (NSF) in Arlington, Virginia to describe the system needed to support an environment conducive to business growth. A space to cultivate ideas and meet with outside investors and customers close to campus is crucial for teams in the early stage of technology development. On-campus programmes to introduce the basics of venture creation for scientists and engineers can help those closest to the bench to evaluate the commercial potential of their work, to better orient themselves to the start-up process and to identify key resources, such as funding, on and off campus.

OUTSIDE HELP

Even funding agencies are waking up to the need to foster an innovation ecosystem to promote the commercialization of academic inventions. Last year, the NSF announced its Innovation Corps, which helps those funded by the foundation to determine the commercial potential of their research. Many funding agencies now ask those applying for grants to explain how their proposed research could be applied to real-world problems. This focus means that graduate students and postdocs have not only the opportunity, but in some cases also the resources, to embark on an entrepreneurial venture.

There is still a long way to go to motivate young scientists to become entrepreneurs. Working in a lab means that graduate students and postdocs are rarely exposed to the practicalities of running a business, and there is still a widespread belief in academia that technology commercialization is a distraction from true scientific research.

Young scientists should make sure that they are in the best position to take advantage of an increasing number of opportunities by keeping in mind the practical applications that may stem from their own research. Seeking out or, better yet, developing their own workshops and seminars on technology commercialization — which is already happening at institutes such as Yale University in New Haven, Connecticut, and the University of California, Berkeley — can help young scientists to understand the practical issues associated with venture creation.

Building a professional network outside the halls of academia, including colleagues in industry and start-up-company investors, will provide early-stage scientists with the crucial resources they will need to draw on, should they find themselves with an entrepreneurial opportunity. Bringing together economic interests and academic pursuits will almost certainly create new challenges. But this can only increase the potential career prospects of a young scientist pursuing a PhD or postdoc. ■

Peter Fiske *is chief executive of PAX Water Technologies in Richmond, California, and author of* Put Your Science to Work *(American Geophysical Union, 2001).*